WMT 2023

Eighth Conference on Machine Translation

December 6–7, 2023

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 979-8-89176-041-7

Introduction

The Eighth Conference on Machine Translation (WMT 2023) took place on Wednesday, December 6 and Thursday, December 7, 2023, immediately preceding the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023).

This is the eighth time WMT has been held as a conference. The first time WMT was held as a conference was at ACL 2016 in Berlin, Germany, the second time at EMNLP 2017 in Copenhagen, Denmark, the third time at EMNLP 2028 in Brussels, Belgium, the fourth time at ACL 2019 in Florence, Italy, the fifth time at EMNLP-2020, which was held as an online event due to the COVID-19 pandemic, the sixth time at EMNLP 2021 at Punta Cana, Dominican Republic, and the seventh time at EMNLP 2022 in Abu Dhabi, United Arab Emirates. Prior to being a conference, WMT was held 10 times as a workshop. WMT was held for the first time at HLT-NAACL 2006 in New York City, USA. In the following years the Workshop on Statistical Machine Translation was held at ACL 2007 in Prague, Czech Republic, ACL 2008, Columbus, Ohio, USA, EACL 2009 in Athens, Greece, ACL 2010 in Uppsala, Sweden, EMNLP 2011 in Edinburgh, Scotland, NAACL 2012 in Montreal, Canada, ACL 2013 in Sofia, Bulgaria, ACL 2014 in Baltimore, USA, EMNLP 2015 in Lisbon, Portugal.

The focus of our conference is to bring together researchers from the area of machine translation and invite selected research papers to be presented at the conference.

Prior to the conference, in addition to soliciting relevant papers for review and possible presentation, we conducted 13 shared tasks. These consisted of 13 translation tasks: General translation, Terminology, Literary translation, Word-level autocompletion, Sign language, Biomedical, Low-resource Indic language translation, Large-scale machine translation evaluation for African languages, Metrics, Quality estimation, MT test suites, Automatic post-editing, and Parallel data curation.

The results of all shared tasks were announced at the conference, and these proceedings also include overview papers for the shared tasks, summarizing the results, as well as providing information about the data used and any procedures that were followed in conducting or scoring the tasks. In addition, there are short papers from each participating team that describe their underlying system in greater detail.

Like in previous years, we have received a far larger number of submissions than we could accept for presentation. WMT 2023 has received 50 full research paper submissions (not counting withdrawn submissions). In total, WMT 2023 featured 18 full research paper presentations and 71 shared task presentations.

WMT 2023 featured a panel on the role of large language models for machine translation. The invited panelists were: Eleftheria Briakou (University of Maryland), Arul Menezes (Microsoft), and José de Souza (Unbabel).

We would like to thank the members of the Program Committee for their timely reviews. We also would like to thank the participants of the shared task and all the other volunteers who helped with the evaluations.

Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz

Co-Organizers

Organizers:

Barry Haddow (University of Edinburgh)
Tom Kocmi (Microsoft)
Philipp Koehn (Johns Hopkins University)
Christof Monz (University of Amsterdam)

Program Committee:

Sadaf Abdul Rauf (Fatima Jinnah Women Unversity)
David Ifeoluwa Adelani (University College London)
Jesujoba Alabi (Saarland University)
Antonios Anastasopoulos (George Mason University)
Mihael Arcan (University of Galway)
Duygu Ataman (New York University)
Eleftherios Avramidis (German Research Center for Artificial Intelligence (DFKI))
Parnia Bahar (AppTek)
Petra Barancikova (Charles University in Prague)
Jasmijn Bastings (Google)
Rachel Bawden (Inria)
Meriem Beloucif (Uppsala University)
Toms Bergmanis (Tilde)
Alexandra Birch (University of Edinburgh)
Frederic Blain (Tilburg University)
Nikolay Bogoychev (University of Edinburgh)
Marine Carpuat (University of Maryland)
Francisco Casacuberta (Universitat Politècnica de València)
Sheila Castilho (Dublin City University)
Boxing Chen (Huawei)
Colin Cherry (Google)
Vishal Chowdhary (MSR)
Chenhui Chu (Kyoto University)
Raj Dabre (NICT)
Steve DeNeefe (RWS Language Weaver)
Michael Denkowski (Amazon)
Shuoyang Ding (Amazon)
Miguel Domingo (Universitat Politècnica de València)
Kevin Duh (Johns Hopkins University)
Koel Dutta Chowdhury (Saarland Informatics Campus, Saarland University)
Hiroshi Echizen'ya (Hokkai-Gakuen University)
Cristina España-Bonet (DFKI GmbH)
Miquel Esplà-Gomis (Universitat d'Alacant)
Mikel L. Forcada (Universitat d'Alacant)
George Foster (Google)
Atsushi Fujita (National Institute of Information and Communications Technology)
Mercedes García-Martínez (Pangeanic)
Jesús González-Rubio (WebInterpret)

Isao Goto (NHK)
Thamme Gowda (Microsoft)
Jeremy Gwinnup (Air Force Research Laboratory)
Thanh-Le Ha (Zoom Video Communications)
Nizar Habash (New York University Abu Dhabi)
Greg Hanneman (Amazon)
Yifan He (Google)
Jindřich Helcl (Charles University in Prague)
John Henderson (Mechanical Learning)
Amr Hendy (Microsoft)
Nico Herbig (German Research Center for Artificial Intelligence (DFKI), Saarland Informatics Campus)
Christian Herold (RWTH Aachen University)
Mika Hämäläinen (Metropolia University of Applied Sciences)
Kenji Imamura (National Institute of Information and Communications Technology)
Yuchen Eleanor Jiang (ETH Zurich)
Josef Jon (Charles University)
Baikjin Jung (POSTECH)
Diptesh Kanojia (University of Surrey)
Huda Khayrallah (Microsoft)
Julia Kreutzer (Google)
Mateusz Krubiński (Charles University)
Roland Kuhn (National Research Council of Canada)
Gaurav Kumar (Bloomberg LP)
Anoop Kunchukuttan (Microsoft AI and Research)
Wen Lai (LMU Munich)
Surafel M. Lakew (Amazon.com, Inc)
Ekaterina Lapshinova-Koltunski (University of Hildesheim)
Samuel Larkin (National Research Council Canada)
Giang Le (University of Illinois Urbana-Champaign)
Gregor Leusch (eBay)
William Lewis (University of Washington)
Samuel Läubli (University of Zurich)
Jean Maillard (Meta AI)
Andreas Maletti (Universität Leipzig)
Arne Mauser (Snowflake)
Antonio Valerio Miceli Barone (The University of Edinburgh)
Amit Moryossef (Bar-Ilan university, University of Zurich)
Kenton Murray (Johns Hopkins University)
Mathias Müller (University of Zurich)
Graeme Nail (University of Edinburgh)
Graham Neubig (Carnegie Mellon University)
Jan Niehues (Karlsruhe Institut of Technology)
Xing Niu (Amazon AI)
Tsuyoshi Okita (Kyushu institute of technology)
Arturo Oncevay (The University of Edinburgh)
Daniel Ortiz-Martínez (University of Barcelona)

Santanu Pal (Wipro)
Jianhui Pang (University of Macau)
Stephan Peitz (Apple)
Sergio Penkale (Lingo24)
Mārcis Pinnis (Tilde)
Maja Popović (ADAPT, Dublin City University)
Reinhard Rapp (University of Mainz)
Vikas Raunak (Microsoft)
Ricardo Rei (Unbabel/INESC-ID)
Matiss Rikters (AIST)
Annette Rios (University of Zurich)
Elizabeth Salesky (Johns Hopkins University)
Rico Sennrich (University of Zurich)
Patrick Simianer (Lilt)
Felix Stahlberg (Google Research)
David Stap (University of Amsterdam)
Katsuhito Sudoh (Nara Institute of Science and Technology (NAIST))
Víctor M. Sánchez-Cartagena (Universitat d'Alacant)
Felipe Sánchez-Martínez (Universitat d'Alacant)
Aleš Tamchyna (Memsource)
Gongbo Tang (Beijing Language and Culture University)
Brian Thompson (Amazon)
Jörg Tiedemann (University of Helsinki)
Antonio Toral (University of Groningen)
Ke Tran (Amazon)
Jonas-Dario Troles (University of Bamberg)
Masao Utiyama (NICT)
David Vilar (Google)
Martin Volk (University of Zurich)
Ekaterina Vylomova (University of Melbourne)
Longyue Wang (Tencent AI Lab)
Wei Wang (Apple AI/ML)
Taro Watanabe (Nara Institute of Science and Technology)
Marion Weller-Di Marco (Ludwig-Maximilians-Universität München)
Tong Xiao (Northeastern University)
François Yvon (ISIR CNRS & Sorbonne Université)
Xianfeng Zeng (Pattern Recognition Center, WeChat AI, Tencent)
Chrysoula Zerva (Instituto de Instituto de Telecomunicações, Instituto Superior Técnico, University of Lisbon)
Dakun Zhang (SYSTRAN)
Zhong Zhou (Carnegie Mellon University)
Vilém Zouhar (ETH Zurich, Charles University)

Table of Contents

Conference Program

Wednesday, December 6, 2023

8:45–9:00 ***Opening Remarks***

9:00–10:30 **Session 1: Shared Task Overview Papers I**

9:00–9:30 *Findings of the 2023 Conference on Machine Translation (WMT23): LLMs Are Here but Not Quite There Yet*
Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović and Mariya Shmatova

9:30–9:45 *Findings of the WMT 2023 Biomedical Translation Shared Task: Evaluation of ChatGPT 3.5 as a Comparison System*
Mariana Neves, Antonio Jimeno Yepes, Aurélie Névéol, Rachel Bawden, Giorgio Maria Di Nunzio, Roland Roller, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Lana Yeganova, Dina Wiemann and Cristian Grozea

9:45–10:00 *Findings of the WMT 2023 Shared Task on Discourse-Level Literary Translation: A Fresh Orb in the Cosmos of LLMs*
Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, Weiyu Chen, Yvette Graham, Bonnie Webber, Philipp Koehn, Andy Way, Yulin Yuan and Shuming Shi

10:00–10:15 *Findings of the Second WMT Shared Task on Sign Language Translation (WMT-SLT23)*
Mathias Müller, Malihe Alikhani, Eleftherios Avramidis, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Sarah Ebling, Cristina España-Bonet, Anne Göhring, Roman Grundkiewicz, Mert Inan, Zifan Jiang, Oscar Koller, Amit Moryossef, Annette Rios, Dimitar Shterionov, Sandra Sidler-Miserez, Katja Tissi and Davy Van Landuyt

10:15–10:30 *Findings of the WMT 2023 Shared Task on Parallel Data Curation*
Steve Sloto, Brian Thompson, Huda Khayrallah, Tobias Domhan, Thamme Gowda and Philipp Koehn

10:30–11:00 ***Coffee Break***

Wednesday, December 6, 2023 (continued)

11:00–12:30 Session 2: Shared Task System Description Posters I

11:00–12:30 General Translation Task

11:00–12:30 *Samsung R&D Institute Philippines at WMT 2023*
Jan Christian Blaise Cruz

11:00–12:30 *NAIST-NICT WMT'23 General MT Task Submission*
Hiroyuki Deguchi, Kenji Imamura, Yuto Nishida, Yusuke Sakai, Justin Vasselli and Taro Watanabe

11:00–12:30 *CUNI at WMT23 General Translation Task: MT and a Genetic Algorithm*
Josef Jon, Martin Popel and Ondřej Bojar

11:00–12:30 *SKIM at WMT 2023 General Translation Task*
Keito Kudo, Takumi Ito, Makoto Morishita and Jun Suzuki

11:00–12:30 *KYB General Machine Translation Systems for WMT23*
Ben LI, Yoko Matsuzaki and Shivam Kalkar

11:00–12:30 *Yishu: Yishu at WMT2023 Translation Task*
Luo Min, yixin tan and Qiulin Chen

11:00–12:30 *PROMT Systems for WMT23 Shared General Translation Task*
Alexander Molchanov and Vladislav Kovalenko

11:00–12:30 *AIST AIRC Submissions to the WMT23 Shared Task*
Matiss Rikters and Makoto Miwa

11:00–12:30 *MUNI-NLP Submission for Czech-Ukrainian Translation Task at WMT23*
Pavel Rychly and Yuliia Teslia

11:00–12:30 *Exploring Prompt Engineering with GPT Language Models for Document-Level Machine Translation: Insights and Findings*
Yangjian Wu and Gang Hu

Wednesday, December 6, 2023 (continued)

11:00–12:30 *Treating General MT Shared Task as a Multi-Domain Adaptation Problem: HW-TSC's Submission to the WMT23 General MT Shared Task*
Zhanglin Wu, Daimeng Wei, Zongyao Li, Zhengzhe YU, Shaojun Li, Xiaoyu Chen, Hengchao Shang, Jiaxin GUO, Yuhao Xie, Lizhi Lei, Hao Yang and Yanfei Jiang

11:00–12:30 *UvA-MT's Participation in the WMT 2023 General Translation Shared Task*
Di Wu, Shaomu Tan, David Stap, Ali Araabi and Christof Monz

11:00–12:30 *Achieving State-of-the-Art Multilingual Translation Model with Minimal Data and Parameters*
Hui Zeng

11:00–12:30 *IOL Research Machine Translation Systems for WMT23 General Machine Translation Shared Task*
Wenbo Zhang

11:00–12:30 *GTCOM and DLUT's Neural Machine Translation Systems for WMT23*
Hao Zong

11:00–12:30 Test Suites

11:00–12:30 *RoCS-MT: Robustness Challenge Set for Machine Translation*
Rachel Bawden and Benoît Sagot

11:00–12:30 *Multifaceted Challenge Set for Evaluating Machine Translation Performance*
Xiaoyu Chen, Daimeng Wei, Zhanglin Wu, Ting Zhu, Hengchao Shang, Zongyao Li, Jiaxin GUO, Ning Xie, Lizhi Lei, Hao Yang and Yanfei Jiang

11:00–12:30 *Linguistically Motivated Evaluation of the 2023 State-of-the-art Machine Translation: Can ChatGPT Outperform NMT?*
Shushen Manakhimova, Eleftherios Avramidis, Vivien Macketanz, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov and Sebastian Möller

11:00–12:30 *IIIT HYD's Submission for WMT23 Test-suite Task*
Ananya Mukherjee and Manish Shrivastava

11:00–12:30 *Test Suites Task: Evaluation of Gender Fairness in MT with MuST-SHE and INES*
Beatrice Savoldi, Marco Gaido, Matteo Negri and Luisa Bentivogli

11:00–12:30 Biomedical Translation Task

11:00–12:30 *Biomedical Parallel Sentence Retrieval Using Large Language Models*
Sheema Firdous and Sadaf Abdul Rauf

11:00–12:30 *The Path to Continuous Domain Adaptation Improvements by HW-TSC for the WMT23 Biomedical Translation Shared Task*
Zhanglin Wu, Daimeng Wei, Zongyao Li, Zhengzhe YU, Shaojun Li, Xiaoyu Chen, Hengchao Shang, Jiaxin GUO, Yuhao Xie, Lizhi Lei, Hao Yang and Yanfei Jiang

11:00–12:30 *Investigating Techniques for a Deeper Understanding of Neural Machine Translation (NMT) Systems through Data Filtering and Fine-tuning Strategies*
Lichao Zhu, Maria Zimina, Maud Bénard, Behnoosh Namdar, Nicolas Ballier, Guillaume Wisniewski and Jean-Baptiste Yunès

11:00–12:30 Literary Translation Task

11:00–12:30 *MAX-ISI System at WMT23 Discourse-Level Literary Translation Task*
Li An, Linghao Jin and Xuezhe Ma

11:00–12:30 *The MAKE-NMTVIZ System Description for the WMT23 Literary Task*
Fabien Lopez, Gabriela González, Damien Hansen, Mariam Nakhle, Behnoosh Namdarzadeh, Nicolas Ballier, Marco Dinarelli, Emmanuelle Esperança-Rodier, Sui He, Sadaf Mohseni, Caroline Rossi, Didier Schwab, Jun Yang, Jean-Baptiste Yunès and Lichao Zhu

11:00–12:30 *DUTNLP System for the WMT2023 Discourse-Level Literary Translation*
Anqi Zhao, Kaiyu Huang, Hao Yu and Degen Huang

11:00–12:30 *HW-TSC's Submissions to the WMT23 Discourse-Level Literary Translation Shared Task*
Yuhao Xie, Zongyao Li, Zhanglin Wu, Daimeng Wei, Xiaoyu Chen, Zhiqiang Rao, Shaojun Li, Hengchao Shang, Jiaxin GUO, Lizhi Lei, Hao Yang and Yanfei Jiang

11:00–12:30 *TJUNLP:System Description for the WMT23 Literary Task in Chinese to English Translation Direction*
Shaolin Zhu and Deyi Xiong

11:00–12:30 African Languages Translation Task

11:00–12:30 *Machine Translation for Nko: Tools, Corpora, and Baseline Results*
Moussa Doumbouya, Baba Mamadi Diané, Solo Farabado Cissé, Djibrila Diané, Abdoulaye Sow, Séré Moussa Doumbouya, Daouda Bangoura, Fodé Moriba Bayo, Ibrahima Sory 2. Conde, Kalo Mory Diané, Chris Piech and Christopher Manning

11:00–12:30 Sign Language Translation Task

11:00–12:30 *TTIC's Submission to WMT-SLT 23*
Marcelo Sandoval-Castaneda, Yanhong Li, Bowen Shi, Diane Brentari, Karen Livescu and Gregory Shakhnarovich

11:00–12:30 *KnowComp Submission for WMT23 Sign Language Translation Task*
Baixuan Xu, Haochen Shi, Tianshi Zheng, Qing Zong, Weiqi Wang, Zhaowei Wang and Yangqiu Song

11:00–12:30 Parallel Data Curation Task

11:00–12:30 *A Fast Method to Filter Noisy Parallel Data WMT2023 Shared Task on Parallel Data Curation*
Nguyen-Hoang Minh-Cong, Nguyen Van Vinh and Nguyen Le-Minh

11:00–12:30 *A Sentence Alignment Approach to Document Alignment and Multi-faceted Filtering for Curating Parallel Sentence Pairs from Web-crawled Data*
Steinthor Steingrimsson

12:30–14:00 ***Lunch Break***

14:00–15:30 Session 3: Research Papers on Document-Level Translation and Use of Large Language Models

14:00–14:15 *Document-Level Language Models for Machine Translation*
Frithjof Petrick, Christian Herold, Pavel Petrushkov, Shahram Khadivi and Hermann Ney

14:15–14:30 *ChatGPT MT: Competitive for High- (but Not Low-) Resource Languages*
Nathaniel Robinson, Perez Ogayo, David R. Mortensen and Graham Neubig

14:30–14:45 *Large Language Models Effectively Leverage Document-level Context for Literary Translation, but Critical Errors Persist*
Marzena Karpinska and Mohit Iyyer

14:45–15:00 *Identifying Context-Dependent Translations for Evaluation Set Production*
Rachel Wicks and Matt Post

15:00–15:15 *Machine Translation with Large Language Models: Prompting, Few-shot Learning, and Fine-tuning with QLoRA*
Xuan Zhang, Navid Rajabi, Kevin Duh and Philipp Koehn

15:15–15:30 *Towards Effective Disambiguation for Machine Translation with Large Language Models*
Vivek Iyer, Pinzhen Chen and Alexandra Birch

15:30–16:00 ***Coffee Break***

Wednesday, December 6, 2023 (continued)

16:00–17:30 Session 4: Research Papers on Translation Modelling

16:00–16:15 *A Closer Look at Transformer Attention for Multilingual Translation*
Jingyi Zhang, Gerard de Melo, Hongfei Xu and Kehai Chen

16:15–16:30 *Bridging the Gap between Position-Based and Content-Based Self-Attention for Neural Machine Translation*
Felix Schmidt and Mattia Di Gangi

16:30–16:45 *Visual Prediction Improves Zero-Shot Cross-Modal Machine Translation*
Tosho Hirasawa, Emanuele Bugliarello, Desmond Elliott and Mamoru Komachi

16:45–17:00 *The Gender-GAP Pipeline: A Gender-Aware Polyglot Pipeline for Gender Characterisation in 55 Languages*
Benjamin Muller, Belen Alastruey, Prangthip Hansanti, Elahe Kalbassi, Christophe Ropers, Eric Smith, Adina Williams, Luke Zettlemoyer, Pierre Andrews and Marta R. Costa-jussà

17:00–17:15 *Towards Better Evaluation for Formality-Controlled English-Japanese Machine Translation*
Edison Marrese-Taylor, Pin Chen Wang and Yutaka Matsuo

17:15–17:30 *There's No Data like Better Data: Using QE Metrics for MT Data Filtering*
Jan-Thorsten Peter, David Vilar, Daniel Deutsch, Mara Finkelstein, Juraj Juraska and Markus Freitag

Thursday, December 7, 2023

9:00–10:30 Session 5: Shared Task Overview Papers II

9:00–9:15 *Results of WMT23 Metrics Shared Task: Metrics Might Be Guilty but References Are Not Innocent*
Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie and George Foster

9:15–9:30 *Findings of the WMT 2023 Shared Task on Quality Estimation*
Frederic Blain, Chrysoula Zerva, Ricardo Ribeiro, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan and André Martins

9:30–9:45 *Findings of the Word-Level AutoCompletion Shared Task in WMT 2023*
Lemao Liu, Francisco Casacuberta, George Foster, Guoping Huang, Philipp Koehn, Geza Kovacs, Shuming Shi, Taro Watanabe and Chengqing Zong

9:45–10:00 *Findings of the WMT 2023 Shared Task on Machine Translation with Terminologies*
Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou and Yuchen Eleanor Jiang

Thursday, December 7, 2023 (continued)

10:00–10:15 *Findings of the WMT 2023 Shared Task on Automatic Post-Editing*
Pushpak Bhattacharyya, Rajen Chatterjee, Markus Freitag, Diptesh Kanojia, Matteo Negri and Marco Turchi

10:15–10:30 *Findings of the WMT 2023 Shared Task on Low-Resource Indic Language Translation*
Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure and Sandeep Kumar Dash

10:30–11:00 ***Coffee Break***

11:00–12:30 **Session 6: Shared Task System Description Posters II**

11:00–12:30 **Metrics Task**

11:00–12:30 *ACES: Translation Accuracy Challenge Sets at WMT 2023*
Chantal Amrhein, Nikita Moghe and Liane Guillou

11:00–12:30 *Challenging the State-of-the-art Machine Translation Metrics from a Linguistic Perspective*
Eleftherios Avramidis, Shushen Manakhimova, Vivien Macketanz and Sebastian Möller

11:00–12:30 *Tokengram_F, a Fast and Accurate Token-based chrF++ Derivative*
Sören DREANO, Derek Molloy and Noel Murphy

11:00–12:30 *Embed_Llama: Using LLM Embeddings for the Metrics Shared Task*
Sören DREANO, Derek Molloy and Noel Murphy

11:00–12:30 *eBLEU: Unexpectedly Good Machine Translation Evaluation Using Simple Word Embeddings*
Muhammad ElNokrashy and Tom Kocmi

11:00–12:30 *Cometoid: Distilling Strong Reference-based Machine Translation Metrics into Even Stronger Quality Estimation Metrics*
Thamme Gowda, Tom Kocmi and Marcin Junczys-Dowmunt

11:00–12:30 *MetricX-23: The Google Submission to the WMT 2023 Metrics Shared Task*
Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh and Markus Freitag

Thursday, December 7, 2023 (continued)

11:00–12:30 *GEMBA-MQM: Detecting Translation Quality Error Spans with GPT-4*
Tom Kocmi and Christian Federmann

11:00–12:30 *Metric Score Landscape Challenge (MSLC23): Understanding Metrics' Performance on a Wider Landscape of Translation Quality*
Chi-kiu Lo, Samuel Larkin and Rebecca Knowles

11:00–12:30 *MEE4 and XLsim : IIIT HYD's Submissions' for WMT23 Metrics Shared Task*
Ananya Mukherjee and Manish Shrivastava

11:00–12:30 *Quality Estimation Using Minimum Bayes Risk*
Subhajit Naskar, Daniel Deutsch and Markus Freitag

11:00–12:30 *Evaluating Metrics for Document-context Evaluation in Machine Translation*
Vikas Raunak, Tom Kocmi and Matt Post

11:00–12:30 *Semantically-Informed Regressive Encoder Score*
Vasiliy Viskov, George Kokush, Daniil Larionov, Steffen Eger and Alexander Panchenko

11:00–12:30 *Empowering a Metric with LLM-assisted Named Entity Annotation: HW-TSC's Submission to the WMT23 Metrics Shared Task*
Zhanglin Wu, Yilun Liu, Min Zhang, Xiaofeng Zhao, Junhao Zhu, Ming Zhu, Xiaosong Qiao, Jingfei Zhang, Ma Miaomiao, Zhao Yanqing, Song Peng, shimin tao, Hao Yang and Yanfei Jiang

11:00–12:30 Quality Estimation Task

11:00–12:30 *Unify Word-level and Span-level Tasks: NJUNLP's Participation for the WMT2023 Quality Estimation Shared Task*
Xiang Geng, Zhejian Lai, Yu Zhang, shimin tao, Hao Yang, Jiajun CHEN and Shujian Huang

11:00–12:30 *HW-TSC 2023 Submission for the Quality Estimation Shared Task*
Yuang Li, Chang Su, Ming Zhu, Mengyao Piao, Xinglin Lyu, Min Zhang and Hao Yang

11:00–12:30 *Scaling up CometKiwi: Unbabel-IST 2023 Submission for the Quality Estimation Shared Task*
Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza and André Martins

11:00–12:30 *SurreyAI 2023 Submission for the Quality Estimation Shared Task*
Archchana Sindhujan, Diptesh Kanojia, Constantin Orasan and Tharindu Ranasinghe

Thursday, December 7, 2023 (continued)

11:00–12:30 *MMT's Submission for the WMT 2023 Quality Estimation Shared Task*
Yulong Wu, Viktor Schlegel, Daniel Beck and Riza Batista-Navarro

11:00–12:30 *IOL Research's Submission for WMT 2023 Quality Estimation Shared Task*
ZEYU YAN

11:00–12:30 Word-Level Autocompletion Task

11:00–12:30 *SJTU-MTLAB's Submission to the WMT23 Word-Level Auto Completion Task*
Xingyu Chen and Rui Wang

11:00–12:30 *PRHLT's Submission to WLAC 2023*
Angel Navarro, Miguel Domingo and Francisco Casacuberta

11:00–12:30 *KnowComp Submission for WMT23 Word-Level AutoCompletion Task*
Yi Wu, Haochen Shi, Weiqi Wang and Yangqiu Song

11:00–12:30 Termninology Translation Task

11:00–12:30 *Terminology-Aware Translation with Constrained Decoding and Large Language Model Prompting*
Nikolay Bogoychev and Pinzhen Chen

11:00–12:30 *Lingua Custodia's Participation at the WMT 2023 Terminology Shared Task*
Jingshu Liu, Mariam Nakhlé, Gaëtan Caillout and Raheel Qadar

11:00–12:30 *Domain Terminology Integration into Machine Translation: Leveraging Large Language Models*
Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, John D. Kelleher, Rejwanul Haque and Andy Way

11:00–12:30 *OPUS-CAT Terminology Systems for the WMT23 Terminology Shared Task*
Tommi Nieminen

11:00–12:30 *VARCO-MT: NCSOFT's WMT'23 Terminology Shared Task Submission*
Geon Woo Park, Junghwa Lee, Meiying Ren, Allison Shindell and Yeonsoo Lee

Thursday, December 7, 2023 (continued)

11:00–12:30 Automatic Postediting Task

11:00–12:30 *HW-TSC's Participation in the WMT 2023 Automatic Post Editing Shared Task*
Jiawei Yu, Min Zhang, Zhao Yanqing, Xiaofeng Zhao, Yuang Li, Su Chang, Yinglu Li, Ma Miaomiao, shimin tao and Hao Yang

11:00–12:30 Indic Languages Translation Task

11:00–12:30 *Neural Machine Translation for English - Manipuri and English - Assamese*
Goutam Agrawal, Rituraj Das, Anupam Biswas and Dalton Meitei Thounaojam

11:00–12:30 *GUIT-NLP's Submission to Shared Task: Low Resource Indic Language Translation*
Mazida Ahmed, Kuwali Talukdar, Parvez Boruah, Prof. Shikhar Kumar Sarma and Kishore Kashyap

11:00–12:30 *NICT-AI4B's Submission to the Indic MT Shared Task in WMT 2023*
Raj Dabre, Jay Gala and Pranjal Chitale

11:00–12:30 *Machine Translation Advancements for Low-Resource Indian Languages in WMT23: CFILT-IITB's Effort for Bridging the Gap*
Pranav Gaikwad, Meet Doshi, Sourabh Deoghare and Pushpak Bhattacharyya

11:00–12:30 *Low-Resource Machine Translation Systems for Indic Languages*
Ivana Kvapilíková and Ondřej Bojar

11:00–12:30 *MUNI-NLP Systems for Low-resource Indic Machine Translation*
Edoardo Signoroni and Pavel Rychly

11:00–12:30 *NITS-CNLP Low-Resource Neural Machine Translation Systems of English-Manipuri Language Pair*
Kshetrimayum Boynao Singh, Avichandra Singh Ningthoujam, Loitongbam Sanayai Meetei, Sivaji Bandyopadhyay and Thoudam Doren Singh

11:00–12:30 *IACS-LRILT: Machine Translation for Low-Resource Indic Languages*
Dhairya Suman, Atanu Mandal, Santanu Pal and Sudip Naskar

11:00–12:30 *IOL Research Machine Translation Systems for WMT23 Low-Resource Indic Language Translation Shared Task*
Wenbo Zhang

Thursday, December 7, 2023 (continued)

14:00–15:30 Session 7: Panel on Large Language Models and Machine Translation

16:00–17:30 Session 8: Research Papers on Evaluation

16:00–16:15 *Trained MT Metrics Learn to Cope with Machine-translated References*
Jannis Vamvas, Tobias Domhan, Sony Trenous, Rico Sennrich and Eva Hasler

16:15–16:30 *Training and Meta-Evaluating Machine Translation Evaluation Metrics at the Paragraph Level*
Daniel Deutsch, Juraj Juraska, Mara Finkelstein and Markus Freitag

16:30–16:45 *Automating Behavioral Testing in Machine Translation*
Javier Ferrando, Matthias Sperber, Hendra Setiawan, Dominic Telaar and Saša Hasan

16:45–17:00 *One Wide Feedforward Is All You Need*
Telmo Pires, António Vilarinho Lopes, Yannick Assogba and Hendra Setiawan

17:00–17:15 *A Benchmark for Evaluating Machine Translation Metrics on Dialects without Standard Orthography*
Noëmi Aepli, Chantal Amrhein, Florian Schottmann and Rico Sennrich

17:15–17:30 *The Devil Is in the Errors: Leveraging Large Language Models for Fine-grained Machine Translation Evaluation*
Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag and Orhan Firat

Findings of the 2023 Conference on Machine Translation (WMT23): LLMs Are Here But Not Quite There Yet

Tom Kocmi Microsoft
Eleftherios Avramidis DFKI
Rachel Bawden Inria, Paris
Ondřej Bojar Charles University
Anton Dvorkovich Dubformer
Christian Federmann Microsoft
Mark Fishel University of Tartu
Markus Freitag Google
Thamme Gowda Microsoft
Roman Grundkiewicz Microsoft
Barry Haddow University of Edinburgh
Philipp Koehn Johns Hopkins University
Benjamin Marie 4i.ai
Christof Monz University of Amsterdam
Makoto Morishita NTT
Kenton Murray Johns Hopkins University
Masaaki Nagata NTT
Toshiaki Nakazawa University of Tokyo
Martin Popel Charles University
Maja Popović Dublin City University
Mariya Shmatova Dubformer
Jun Suzuki Tohoku University

Abstract

This paper presents the results of the General Machine Translation Task organised as part of the 2023 Conference on Machine Translation (WMT). In the general MT task, participants were asked to build machine translation systems for any of 8 language pairs (corresponding to 14 translation directions), to be evaluated on test sets consisting of up to four different domains. We evaluate system outputs with professional human annotators using a combination of source-based Direct Assessment and scalar quality metric (DA+SQM).

1 Introduction

The Eighth Conference on Machine Translation (WMT23)[1] was held at EMNLP 2023 and hosted a number of shared tasks on various aspects of machine translation (MT). This conference built on 17 previous editions of WMT as a workshop or a conference (Koehn and Monz, 2006; Callison-Burch et al., 2007, 2008, 2009, 2010, 2011, 2012; Bojar et al., 2013, 2014, 2015, 2016, 2017, 2018; Barrault et al., 2019, 2020; Akhbardeh et al., 2021; Kocmi et al., 2022).

Following last year's shift from focusing mainly on the news domain, we have continued to explore the capabilities of "General Machine Translation". While the news domain provided a clear and familiar benchmark, we realized the need to test MT in more diverse settings. Our goal is to assess MT systems' ability to handle a broader range of language use. How to test general MT performance is a research question in itself. Countless phenomena could be evaluated, the most important being:

- various domains (news, medicine, IT, patents, legal, social, gaming, etc.)
- style of text (formal or spoken language, fiction, technical reports, etc.)
- robustness to non-standard (or noisy) user-generated content (grammatical errors, code-switching, abbreviations, etc.)

Evaluating all phenomena is nearly impossible and creates numerous unforeseen problems. Therefore, we decided to simplify the problem and start with an evaluation of different domains. We selected the following domains: news, e-commerce, social/user-generated content (UGC), speech, and manuals. They were chosen to represent topics with different content styles and to be understandable for humans without special in-domain knowledge, thus not requiring specialized translators or human raters for evaluation. Due to limited access

[1] http://www2.statmt.org/wmt23/

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 1–42
December 6–7, 2023.

to monolingual data across all languages, each language direction contains only a subset of up to four domains.

In addition to language pairs evaluated last year:

Czech→Ukrainian,
English↔Chinese,
English→Czech,
English↔German,
English↔Japanese,
English↔Russian,
Ukrainian→English,

we introduce a new language pair to WMT, namely:

English↔Hebrew.

Other than language pairs, there are several differences with respect to last year's task. All language pairs are provided with the sentence boundaries marked except for English↔German, where we decided to experiment with paragraph-level translation. Another significant change for this year is the unification of our human evaluation protocol. We no longer rely on reference-based MTurk evaluation and move the evaluation towards source-based DA+SQM evaluation (introduced last year) with professional annotators. Finally, this year's shared task included an increased number of test suites (Section 6), allowing the evaluation of MT outputs from different perspectives, including a range of linguistic phenomena, purposely difficult sentences, specialist domains, gendered translations and non-standard UGC translation.

All General MT task submissions, sources, references and human judgements are available at Github [2]. The interactive visualization and comparison of differences between systems can be browsed online on an interactive leaderboard[3] using MT-ComparEval (Klejch et al., 2015; Sudarikov et al., 2016).

The structure of the paper is as follows. We describe the process of collecting, cleaning and translating the test sets in Section 2 followed by a summary of the permitted training data for the constrained track Section 3. We list all submitted systems in Section 4. The human evaluation approach of DA+SQM is described in Section 5. Finally, Section 6 describes the test suites and summarises their conclusions.

[2] https://github.com/wmt-conference/wmt23-news-systems

[3] http://wmt.ufal.cz

Summary of the WMT2023 General MT task
The main findings are as follows:

- Large Language Models (LLMs) exhibit strong performance across the majority of language pairs, although this is based only on two LLM-based system submissions. Test suite analysis revealed that although GPT4 excelled in some areas (e.g. UGC translation) struggled with other aspects such as speaker gender translation and specific domains (e.g. legal), whereas it ranked lower than encoder-decoder systems when translating from English into less-represented languages (e.g. Czech and Russian)

- We have observed a decline in the number of submissions into the constrained track. Consequently, we plan to re-evaluate the definition and the incentives of the constrained track and consider incorporating open-source LLMs in future evaluations.

- We demonstrate the feasibility of paragraph-level German↔English tasks, although more investigation would be required before generalising to all language pairs.

- Professional human translations do not always guarantee high quality. For Hebrew↔English, our references are likely to be post-edited MT, while for Chinese→English, the reference translation is worse than the majority of automatic translations.

- The manual evaluation results obtained from DA+SQM and MQM methods yield comparable cluster rankings.

2 Test Data

In this section, we describe the process of collecting data in Section 2.1, followed by the explanation of preprocessing steps in Section 2.2. Producing human references is summarized in Section 2.3 and lastly test set analysis is conducted in Section 2.4.

2.1 Collecting test data

As in the previous years, the test sets consist of unseen translations collected especially for the task. This has become even more important with the rise of LLMs trained on unspecified training data. To prevent possible contamination, we focused on collecting as recent data as possible across various

Lang. pair	Domain name	Domain type	#docs	#segs	#segs/#docs
cs→uk	*	*	156	2017	12.93
	games	News	17	180	10.59
	news	News	35	567	16.20
	official	Social/UGC	26	347	13.35
	personal	Social/UGC	31	390	12.58
	voice	Speech	47	533	11.34
de→en	*	*	210	549	2.61
	manuals	Manuals	15	74	4.93
	mastodon	Social/UGC	95	103	1.08
	news	News	47	277	5.89
	user_review	E-commerce	53	95	1.79
en→{cs,he,ja,ru,uk,zh}	*	*	192	2074	10.80
	mastodon	Social/UGC	79	504	6.38
	news	News	30	516	17.20
	speech	Meeting notes	25	547	21.88
	user_review	E-commerce	58	507	8.74
en→de	*	*	192	557	2.90
	mastodon	Social/UGC	79	212	2.68
	news	News	30	139	4.63
	speech	Meeting notes	25	113	4.52
	user_review	E-commerce	58	93	1.60
he→en	*	*	94	1910	20.32
	news	News	68	1558	22.91
	reviews	Social/UGC	26	352	13.54
ja→en	*	*	282	1992	7.06
	ad	Social/UGC	53	245	4.62
	ec	Social/UGC	25	255	10.20
	news	News	37	495	13.38
	qa	Conversational	118	497	4.21
	user_review	E-commerce	49	500	10.20
ru→en	*	*	162	1723	10.64
	manuals	Manuals	15	505	33.67
	news	News	54	676	12.52
	reviews	Social/UGC	93	542	5.83
uk→en	*	*	132	1826	13.83
	clipboard	Social/UGC	30	504	16.80
	news	News	26	514	19.77
	other	Social/UGC	27	538	19.93
	voice	Speech	49	270	5.51
zh→en	*	*	179	1976	11.04
	manuals	Manuals	14	487	34.79
	news	News	38	763	20.08
	user_review	E-commerce	127	726	5.72

Table 1: Test set statistics per direction and domain (rows marked * are over all domains). Note that en→de shares source test data with the other from-English directions, but as translation and evaluation for both en→de and de→en were carried out on the paragraph level (a segment therefore being a paragraph rather than a sentence), this results in a lower number of segments per document. The domain name is as indicated in the released test sets and domain type indicates the broader domain category.

domains. This task is incredibly difficult and needs further investigation in future years. There are three main limitations:

- Finding sources with different domains.
- Finding data that are in the public domain or under open licenses.
- Finding recently created data to minimize the risk of them being part of the training pipelines.

The test sets are publicly released to be used as translation benchmarks. Here we describe the test sets' production and composition.

We decided to collect data from 5 domains (news, social/user-generated, e-commerce, manuals, and speech). For all language pairs, we aimed for a test set size of 2,000 sentences and to ensure that the test sets were "source-original", namely that the source text was first written in the source language, and then the target text is the human

translation. This is to avoid "translationese" effects on the source language, which can have a detrimental impact on the accuracy of evaluation (Toral et al., 2018; Freitag et al., 2019; Läubli et al., 2020; Graham et al., 2020). We collected roughly the same number of sentences for each domain. For some languages, we could not locate high quality data and therefore we selected more sentences from other domains. Note that descriptions in this section refer to source monolingual data when mentioning a language.

News domain For most languages this domain contains data prepared in the same way as in previous years (Akhbardeh et al., 2021). We collected news articles from February 2023 extracted from online news sites, preserving document boundaries. We expect that news domain text will generally be of high quality. The news in Hebrew was kindly provided by the Israeli Association of Human Language Technologies (IAHLT).[4] These are samples of originally Hebrew texts from news published in Israel Hayom[5] in 2022.

E-commerce domain (product reviews) This domain consists of user reviews of different Amazon products selected from the publicly available multilingual corpus (Keung et al., 2020). This corpus was designed for multilingual text classification and consists of reviews written in English, Japanese, German, French, Spanish, and Chinese, between 2015 and 2019. We used the test parts of the English, German, Japanese and Chinese corpora for extracting the source part of the WMT test set. The reviews were selected so that the resulting corpus covers each product, all rating scores for the product, and the lexical diversity is maximized. The lexical diversity was estimated as a simple ratio between the number of distinct words/characters (vocabulary) divided by the total number of words/characters.

Social/user-generated domain For English and German, we relied on the Mastodon Social API.[6] Mastodon is a federated social network that is compatible with the W3C standard ActivityPub (Webber et al., 2018). Users publish short-form content similar to tweets that are referred to as "toots" for historical reasons. As this is a decentralized social media network, different servers have very different data, policies, communities, and uses. We decided to use `mastodon.social`, the original server, as it has a large community as well as publicly available toots. We collected data in early May of 2023. We used the reported language ID label, but were only able to collect enough data in German and English. We only collected toots with more than 150 characters in length in order to allow for data that was more likely to be semantically interesting for evaluating translation systems.

For Hebrew, we used comments on news articles from the Israel Hayom site mentioned above. This data was also provided by IAHLT.

For Russian, we used data from the Geo Reviews Dataset containing reviews about organizations published on Yandex Maps and open for academic and research purposes.[7]

For Japanese, we used product descriptions of a b2b e-commerce site and search advertising text ads for the social and user-generated domain, because we could not obtain high-quality data for this domain type. MonotaRo Co., Ltd. provided product descriptions of their private label brands listed on their b2b e-commerce site.[8] We defined a document for a product description as a combination of a title, product description, and cautionary note. CyberAgent, Inc.[9] provided search advertising text ads with their client's consent. We defined a document for an ad as the longest possible combination of multiple titles and descriptions.

Manuals For this domain, we primarily sourced scanned versions of different mostly gaming manuals provided by Centific[10]. These were then converted to digital text format using Optical Character Recognition (OCR) technology. Given the inaccuracies of OCR, the digitized content underwent a subsequent post-editing phase, where humans reviewed and corrected any errors. The selection of manuals ranged across various sources, and none of them were older than five years.

Speech The exact data types used in the "conversational" or "speech" domain vary across language pairs.

For English→Czech, the data comes from the test set which was created for the 2023 instance of

[4] https://www.iahlt.org

[5] https://www.israelhayom.co.il

[6] https://mastodon.social/api/v1/timelines/public

[7] https://github.com/yandex/geo-reviews-dataset-2023

[8] https://www.monotaro.com/

[9] https://www.cyberagent.co.jp

[10] https://www.centific.com

AutoMin 2023 (Ghosal et al., 2022).[11] The texts are manually curated transcripts of project meetings, same in style as released in ELITR Minuting Corpus (Nedoluzhko et al., 2022). The meetings were held mostly remotely or in a hybrid form, all meeting participants were non-native speakers of English and the meetings were always on rather technical and in-depth topics. Our manual curation corrected ASR errors (but not errors in English grammar or vocabulary) and de-identified the transcripts, replacing names with placeholders ("PERSONxy", "PROJECTxy" and similar). For person names, round brackets are used at the beginnings of lines to indicate the speaker and square brackets are used in the text when the person was mentioned. The data contain also some markup, e.g. "<unintelligible/>". These conventions are likely to be distorted by translation systems and we also noticed that they were distorted in the reference translation (the style of the brackets was ignored). This tiny detail can influence both manual and automatic scoring on this domain.

For Japanese, we used question-answer pairs from a community question-answering service. NTT Resonant Inc., which recently merged with NTT DOCOMO, INC., provided question-answer pairs from their website, *Oshiete! goo*.[12] For every question-answer pair, we defined a document as a combination of a question and its best answer marked by the user.

Czech and Ukrainian source texts Source texts for Czech→Ukrainian and Ukrainian→English translation included the News domain as described above and texts collected through the Charles Translator for Ukraine.[13] With users' consent, the service can log their inputs for the purpose of creating a dataset of real use cases. The datasets are extracted from the inputs collected from May 2022 to April 2023.

The Charles Translator mobile app supports voice input, which is converted to text using Google ASR (automatic speech recognition). The texts collected this way were marked as the `voice` domain. For Ukrainian→English, the remaining Ukrainian inputs were classified either as `clipboard` (texts inserted to the Charles Translator using the *Paste from clipboard* button) and `other`. The `clipboard` texts are more likely to include formal communication copied from web sites, but we noticed it includes personal communication (copied from chat applications) as well. Thus for Czech→Ukrainian, we decided to classify the remaining Czech inputs either as `official` (formal communication) or `personal` (personal communication), ignoring whether they were inserted from a clipboard or written using a keyboard.

The texts were filtered and pseudonymized in the same way as last year (Kocmi et al., 2022), so for example we asked the annotators not to delete or fix noisy inputs as long as they are comprehensible. There was one exception from this rule this year: the Czech `voice` domain data was post-edited to fix ASR errors, including missing punctuation and casing.

The source texts were translated by professional translators principally following the brief in Appendix C. Last year, parts of the Ukrainian→Czech test set was detected to be post-edited MT. Therefore this year, we decided to hire two professional translators directly without the mediation of a translation agency, we emphasised the rule that the translations must be done from scratch (without MT postediting and without translation memories). We could not detect any MT postediting in the resulting translations.

2.2 Human preprocessing of test data

Although testing of robustness of MT is an important task, the noisy data introduces problems for human translators and annotators. Therefore, we decided to discard data considered too noisy. Furthermore, publicly available data often contains inappropriate content, which can stress either human translators or human annotators, leading to a decrease in the quality (for example, translators refuse to translate political content considered censored in their countries).

Therefore, we asked humans to check collected data and carry out minor corrections (mainly checking sentence splits and discarding similar or repeated content). This was sufficient for the news domain because it was often clean and without serious problems. However, with the expansion towards general MT, we find ourselves running into an issue of source data being noisier and less well formatted and that therefore needs to be handled before translation. Furthermore, we asked them to remove shortest documents to keep longer context. The source data for test sets therefore goes through

[11] https://ufal.github.io/automin-2023/
[12] https://oshiete.goo.ne.jp/
[13] http://translator.cuni.cz

human validation checks involving linguists discarding inappropriate content altogether and carrying out minor textual corrections to the data. You can find the linguistic brief for prepossessing in Appendix B.

2.3 Test set translation

The translation of the test sets was performed by professional translation agencies, according to the brief in Appendix C. Different partners sponsored each language pair and various translation agencies were therefore used, which may affect the quality of the translation.

Regrettably, upon reviewing translations procured from one of the agencies (the one responsible for English to Hebrew and Hebrew to English translations), it appeared that the translations might have been post-edited from publicly available online translation systems. This observation contradicts the initial instruction provided for agency that precluded the use of any automated translation platforms. While the agency has asserted that their professional translations conducted translations from scratch, our evaluation suggested otherwise. Moving forward, we propose to build a step-by-step verification system to avoid such discrepancies.

Human translations would not be possible without the sponsorship of our partners: Microsoft, Toloka AI, Google, Charles University, NTT, and Dubformer.

2.4 Test set analysis

As described previously, the chosen domains, sources for the data and the number of sentences per domain was subject to the availability of high quality data in each language direction. For example, while the news domain was available for all language directions, social media data was only available for English, German (both from Mastodon) and Hebrew (from comments on news articles). The number of documents, segments, average document length and type-token ratio (of the source side of the test sets) are given in Table 1.

Document context Document context is available for all language directions, although the average document length varies both by domain and language direction. Manuals tend to represent the longest domains, followed by the news domain. The social media domain tends to represent the shortest documents. along with reviews. Note that this year, we piloted translation and evaluation of en→de and de→en at the paragraph level (with each segment therefore containing several sentences), with the aim of avoiding the constraint of having a one-to-one mapping at the level of the sentence between source texts and their translations. This is visible in the statistics in Table 1 as the number of segments is lower for these two directions, as is the average document length.

Lexical diversity We can compare the type-token ratio (TTR) to get an idea of the relative lexical diversity of (i) domains and (ii) original vs. translated sentences.[14,15] Raw TTRs for each language pair and domain are shown in Table 11 in Appendix D. Regarding domains, the TTR appears highest for texts mastodon, perhaps illustrating the diversity of conversational topics and also of the potentially non-standard nature of the texts. User reviews appear to have the lowest TTR, most likely due to the fact that similar vocabulary is used across reviews. The TTR of course differs according to the language in question, according to the differing morphological properties.

Anonymisation and markup A particularity of the 'speech' domain is the presence of placeholders for anonymised elements and markup (in the form of tags). For example, there are 35 placeholders surrounded either by square or rounded brackets to indicate different people, organisations and projects (e.g. (PERSON1), [PERSON9], [ORGANIZATION4], [PROJECT8], etc.). The 'person' tags are used both in-text to replace the names of people and at the beginning of lines to indicate who is talking. Markup is added to indicate speakers talking at the same time (<parallel_talk>), unintelligible passages (<unintelligible/>), laughter (<laugh/>) and other noise (<other_noise/>).

2.5 Test suites

In addition to the test sets of the regular domains, the test sets given to the system participants were augmented with several *test suites*, i.e. custom-made test sets focusing on particular aspects of MT translation. The test suites were contributed and evaluated by test suite providers as part of a

[14]The TTR is the ratio of unique tokens to total tokens, and it is higher the diverse the vocabulary of a text is. It is dependent on the morphological complexity of a language, but can also vary due to other factors.

[15]Texts are tokenised using the language-specific Spacy models (Honnibal and Montani, 2017) where available. For Hebrew, we took the multilingual Spacy model, since a language-specific one was not available.

decentralized sub-task, which will be detailed in Section 6.

3 Training Data

Similar to the previous years, we provide a selection of parallel and monolingual corpora for model training. The provenance and statistics of the selected parallel datasets are provided in Appendix in Table 9 and Table 10. Specifically, our parallel data selection include large multilingual corpora such as Europarl-v10 (Koehn, 2005), Paracrawl-v9 (Bañón et al., 2020), CommonCrawl, NewsCommentary-v18, WikiTitles-v3, WikiMatrix (Schwenk et al., 2021), TildeCorpus (Rozis and Skadiņš, 2017), OPUS (Tiedemann, 2012), UN Parallel Corpus (Ziemski et al., 2016), and language-specific corpora such as CzEng-v2.0 (Kocmi et al., 2020), YandexCorpus,[16] ELRC EU Acts, JParaCrawl (Morishita et al., 2020), Japanese-English Subtitle Corpus (Pryzant et al., 2018), KFTT(Neubig, 2011), TED (Cettolo et al., 2012), CCMT, and back-translated news. Links for downloading these datasets were provided on the task web page;[17] in addition, we automated the data preparation pipeline using MTDATA (Gowda et al., 2021).[18] MTDATA downloads all the mentioned datasets, except CCMT and CzEng-v2.0, which required user authentication. This year's monolingual data include the following: News Crawl, News Discussions, News Commentary, CommonCrawl, Europarl-v10 (Koehn, 2005), Extended CommonCrawl (Conneau et al., 2020), Leipzig Corpora (Goldhahn et al., 2012), UberText and Legal Ukrainian.

4 System submissions

This year, we received a total of 72 primary submissions from 17 participants. In addition, we collected translations from online MT systems across all language pairs. Online system outputs come from 6 public MT services and were anonymized as ONLINE-{A,B,G,M,W,Y}, which added additional 77 system outputs. The participating systems are listed in Table 2 and detailed in the rest of this section.

Finally, we added translations by three contrastive systems. Two of them are based on the NLLB translation model (NLLB Team et al., 2022) modified by (Freitag et al., 2023) to have a suboptimal performance, using (i) greedy search (NLLB_Greedy) and (ii) following minimum Bayes risk decoding (MBR) optimizing the BLEU metric (NLLB_MBR_BLEU). Neither of them is the official (and better performing) NLLB model. The third contrastive translation is produced by the large language model GPT4 using 5-shot prompting with fixed random translation examples, using the exact prompt by Hendy et al. (2023) together with their predefined few-shot examples. For languages not evaluated in their study, we took examples from the last WMT test sets.

Appendix E provides details of the submitted systems if the authors provided such details.

4.1 Constrained and unconstrained tracks

For presentation of the results, systems are treated as either constrained or unconstrained. A system is classified as constrained if the authors reported training only on the provided data and adhering to the rules describing the use of publicly available pre-trained models. The constrained track imposes restrictions on training data, metrics, and pretrained models, while the unconstrained track provides unrestrained flexibility.

The constrained track limitations are mainly around the training and testing data, together with the limitation on pretrained models:

- **Training data:** Only data specified for the current year are permissible, see Section 3. Multilingual systems can be used as long as they only use WMT23 data.
- **Metrics:** The training pipeline can use pretrained metrics evaluated in previous WMT Metrics shared tasks, e.g., COMET (Rei et al., 2022), Bleurt (Yan et al., 2023).
- **Pretrained models:** only the following list of models is allowed together with all their public sizes: mBART (Liu et al., 2020), BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLM-RoBERTa (Conneau et al., 2020), sBERT (Reimers and Gurevych, 2019), and LaBSE (Feng et al., 2022).
- **Linguistic tools:** Basic tools like taggers, parsers, and morphology analyzers are allowed.

[16] https://github.com/mashashma/WMT2022-data
[17] https://statmt.org/wmt23/translation-task.html
[18] http://www2.statmt.org/wmt23/mtdata

Submission Name	Language Pairs	System Description
AIRC	de-en, en-ja, ja-en, en-de	(Rikters and Miwa, 2023)
ANVITA	ja-en, zh-en, en-ja, en-zh	(no associated paper)
CUNI-DocTransformer	en-cs	(Popel, 2020)
CUNI-GA	en-cs, cs-uk	(Jon et al., 2023)
CUNI-Transformer	en-cs, cs-uk	(Popel, 2020)
GPT4-5shot	**All language pairs**	(Hendy et al., 2023)
GTCOM	de-en, ja-en, he-en, en-cs, en-he, cs-uk, en-uk, uk-en	(Zong, 2023)
HW-TSC	de-en, en-zh, zh-en	(Wu et al., 2023b)
IOL-Research	zh-en, en-zh	(Zhang, 2023)
TeamKYB	ja-en, en-ja	(LI et al., 2023)
Lan-BridgeMT	**All language pairs**	(Wu and Hu, 2023)
MUNI-NLP	cs-uk	(Rychlý and Teslia, 2023)
NAIST-NICT	en-ja, ja-en	(Deguchi et al., 2023)
NLLB_Greedy	**All language pairs**	(Freitag et al., 2023)
NLLB_MBR_BLEU	**All language pairs**	(Freitag et al., 2023)
ONLINE-A	**All language pairs**	-
ONLINE-B	**All language pairs**	-
ONLINE-G	**All language pairs**	-
ONLINE-M	en-ru, zh-en, en-zh, de-en, en-cs, ja-en, en-de, en-ja, ru-en	-
ONLINE-W	en-uk, ja-en, de-en, en-ja, ru-en, en-de, uk-en, en-ru, zh-en, en-cs, en-zh, cs-uk	-
ONLINE-Y	**All language pairs**	-
PROMT	en-ru, ru-en	(Molchanov and Kovalenko, 2023)
SRPH	he-en, en-he	(Cruz, 2023)
SKIM	en-ja, ja-en	(Kudo et al., 2023)
UPCite-CLILLF	fr-en, en-fr	(no associated paper)
UvA-LTL	he-en, en-he	(Wu et al., 2023a)
YiShu	zh-en, en-zh	(Min et al., 2023)
LanguageX	en-zh, en-uk, ru-en, uk-en, en-de, he-en, ja-en, zh-en, en-he, de-en, en-cs, en-ja, en-ru	(Zeng, 2023)

Table 2: Participants in the General MT shared task. Online system translations were not submitted by their respective companies but were obtained by us, and are therefore anonymized in a fashion consistent with previous editions of the task.

The online systems and contrastive systems are treated as unconstrained during the automatic and human evaluation.

4.2 OCELoT

We used the open-source OCELoT platform[19] to collect system submissions again this year. The platform provides anonymized public leaderboards[20] and was also used for two other WMT23 shared tasks: Biomedical (Neves et al., 2023) and Sign Language Translation (Müller et al., 2023). As in previous years, only registered and verified teams with correct contact information were allowed to submit their system outputs and each verified team was limited to 7 submissions per test set. Submissions on leaderboards with BLEU (Papineni et al., 2002) and CHRF (Popović, 2015) scores from SacreBLEU (Post, 2018) were displayed anonymously to avoid publishing rankings based on automatic scores during the submission period. Until one week after the submission period, teams could select a single primary submission per test set, specify if the primary submission followed a constrained or unconstrained setting, and submit a system description paper abstract. These were mandatory for a system submission to be included in the human evaluation campaign.

5 Human Evaluation

Human evaluation for all language translation directions is performed with source-based ("bilingual") Direct Assessment (DA, Graham et al., 2013) of individual segments in document context with Scalar Quality Metrics (SQM) guidelines, mostly following the setup established at WMT22 (DA+SQM, Kocmi et al., 2022). DA+SQM asks the annotators to provide a score between 0 and 100 on a sliding scale, but the slider is presented with seven labelled tick marks, as demonstrated in Figure 1.

Two different annotation platforms and four distinct pools of annotators (Table 3) are used for annotation of different language pairs. We use the open-source framework Appraise (Federmann, 2018) for the evaluation of English→Czech, English↔{Chinese, German, Japanese}, and Czech→Ukrainian. Toloka AI[21] hosts the evaluation of English↔{Hebrew, Russian, Ukrainian} using their own implementation of the source-based document-level DA+SQM task, which is as close as possible to the Appraise user interface.

We keep the selection process of documents for annotation mostly the same as in the previous year. The only change made in order to align closer with the MQM-based evaluation run at the Metrics shared task (Freitag et al., 2023) is to present the first 10 segments from a document instead of random 10 consecutive segments.

We again collect both segment-level scores and document-level scores, but compute rankings based on segment scores only.

5.1 Human annotators

Annotations for different language pairs are provided by four different parties with their pool of annotators of distinct profiles as presented in Table 3. We shift towards more professional or semi-professional annotators' pools and decide not to use MTurk annotations as in past years for reference-based DA evaluation for into-English language directions.

Assessments for English↔{Chinese, German, Japanese} are provided by Microsoft and their pool of bilingual target-language native speakers, professional translators or linguists, highly experienced in MT evaluation. Microsoft monitors the annotators' performance over time and permanently removes from the pool those who fail quality control, which increases the overall quality of the human assessment.

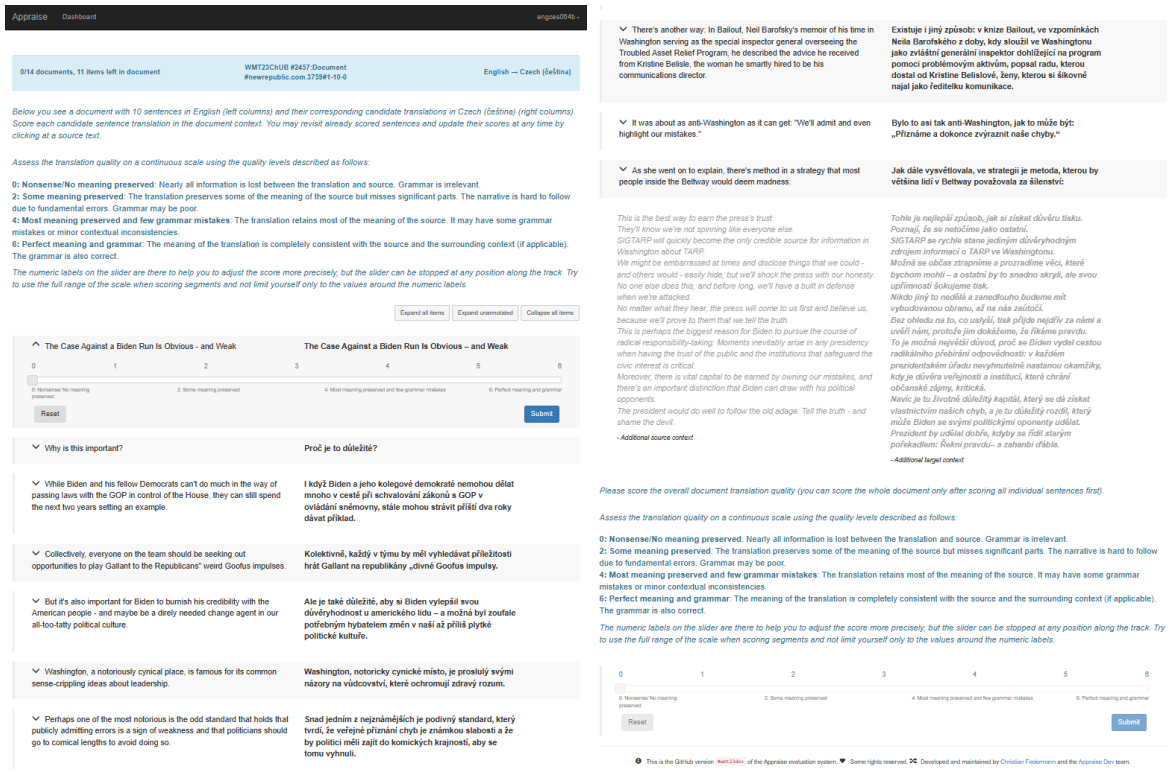
Charles University provides annotators for language pairs involving the Czech language, i.e., English→Czech and Czech→Ukrainian. Their annotators are linguists, translators, researchers and students who are native speakers of the target language with high proficiency in the source language.

DA scores for English↔{Hebrew, Russian, Ukrainian} are collected by Toloka AI using their paid crowd of bilingual target-language native speakers. Toloka AI tests proficiency of their annotator crowd across different NLP annotation tasks and allowed only annotators who deemed reliable according to their quality control measures.

5.2 Document selection and quality control

The document selection process remains the same as in the previous year with minor changes. We first randomly sample a subset of document snippets from each of the domains for annotations, sampling the domains with approximately the same number of segments per domain. This ensures that

[19] https://github.com/AppraiseDev/OCELoT
[20] https://ocelot-wmt23.mteval.org
[21] https://toloka.ai



(a) Top part of the screen with segment-level scoring.

(b) Bottom part of the screen with document-level scoring.

Figure 1: Screenshot of the document-level DA+SQM configuration in the Appraise interface for an example assessment from the human evaluation campaign for out of English language pairs. The annotator is presented with the entire translated document snippet randomly selected from competing systems (anonymized) with additional static contexts, and is asked to rate the translation of individual segments and then the entire document on sliding scales between 0 and 100.

all systems in the given language pairs are evaluated on the same subset of the test set, allowing fair comparison between them. As in previous years, we aim to collect approximately 1,500 assessments per system per language pair. Due to concerns about having sufficient annotations, we create two batches of HITs, each providing half of the required assessments, such that at least all segments in the first batch could be covered for all systems, with the second campaign completed if possible.

For HIT generation for English↔German, which feature paragraph-level test sets (documents consist of paragraphs instead of sentences), we simply consider a whole paragraph as a "segment", collecting paragraph-level assessments. In that regard, we collect fewer DA scores per system comparing to other language pairs, but the human evaluation covers a larger subset of the testsets.

Last year, we used snippets of at most 10 randomly selected consecutive segments from a document as "documents" for document-level annotation. This year, we use 10 first segments from a document instead, in order to align with the MQM-based evaluation used at the Metrics shared task (Freitag et al., 2023).

All HITs consist of exactly 100 segments and are generated as in the past:

1. Snippet-system pairs are randomly sampled (from the restricted set of pre-sampled snippets) to create up to 80 segments;
2. Random snippets for the remaining 20 (or more) segments are duplicated from the first 80 to serve as quality control items;
3. BAD references are introduced to the random segments in the duplicated snippets to have about 12-14% of quality control segments per HIT.

BAD translations are created by replacing an embedded sequences of tokens in the segment with a random phrase of the same length from a different reference segment.[22]

We perform quality control by measuring an annotator's ability to reliably score BAD translations

[22]For full details, see the HIT and batch generation code: https://github.com/wmt-conference/wmt23-news-systems

Language pairs	Annotators' profile	Tool
English↔Chinese/German/Japanese	Microsoft annotators: bilingual target-language native speakers, professional translators or linguists, experienced in MT evaluation	Appraise
Czech→Ukrainian	Paid translators and target-language native speakers	Appraise
English→Czech	Czech paid linguists, annotators, researchers, students with high proficiency in English	Appraise
English↔Hebrew/Russian/Ukrainian	Toloka AI paid crowd: bilingual target-language native speakers high-performing in other task types	Toloka.ai

Table 3: Annotators' profiles and annotation tools for each language pair in human evaluation.

Language Pair	Sys.	Assess.	Assess/Sys
Chinese→English	16	20,535	1283.4
Czech→Ukrainian	14	23,191	1656.5
German→English	14	13,573	969.5
English→Chinese	16	24,551	1534.4
English→Czech	16	25,527	1595.4
English→German	13	14,267	1097.5
English→Japanese	17	26,115	1536.2
Japanese→English	18	27,858	1547.7

Table 4: Amount of segments evaluated in the WMT23 manual evaluation campaign; including human references as systems; after excluding quality control items and document-level scores.

Language Pair	Ann.	HITs	HITs/Ann.
Chinese→English	13	128	9.8
Czech→Ukrainian	9	146	16.2
German→English	21	82	3.9
English→Czech	36	162	4.5
English→German	22	87	4.0
English→Japanese	21	164	7.8
English→Chinese	13	154	11.8
Japanese→English	20	174	8.7

Table 5: Numbers of individual annotators taking part in the WMT23 human evaluation campaign and the average number of HITs collected per annotator.

significantly lower than corresponding original system outputs using a paired significance test with $p < 0.05$. We pair two HITs into a single annotation task with about 24-28 quality control segments to ensure a sufficient sample size for the statistical test. In campaigns hosted on Appraise, if an annotator is not able to demonstrate reliability on BAD references, they are excluded from further annotations, the HITs are reset and annotated from scratch by another annotator if possible.

The total number of assessments collected for each language pair and the average number of assessments per system in WMT23 manual evaluation are presented in Table 4.

5.3 Calibration HITs

Last year we introduced calibration HITs, which this year we collect for all language pairs. A calibration HIT is a HIT with 100 randomly selected segments, which is identical for and completed by all annotators, in addition to their regular annotation HITs. We release these alongside the other annotations and the anonymized mapping between annotators and HITs in order to enable additional analysis. With a small set of sentences annotated by all annotators, we are better able to examine questions about inter-annotator consistency and provide data for future research in this area.

Table 5 shows the number of unique annotators per language pair along with the total number of HITs and average number of HITs per annotator. We leave more detailed analysis of collected calibration data to future work.

5.4 Human ranking computation

The official rankings shown in Table 6 are generated on the basis of the segment-level raw DA+SQM scores that are collected within document context for all language pairs.[23] Whole documents with at least one quality control segment (i.e., BAD references) and HITs that failed to pass quality control are removed prior to computing the rankings.[24]

In this year's evaluation, we have chosen not to normalize scores by discontinuing the use of z-scores, given their potential to exacerbate system comparisons (Knowles, 2021). While utilizing raw scores is not flawless—considering each annotator employs distinct annotation strategies — we have sought to counteract this by distributing

[23]The code used to generate the rankings in Table 6 can be found here: https://github.com/AppraiseDev/Appraise/blob/main/Campaign/management/commands/ComputeWMT23Results.py

[24]Two HITs for Czech→Ukrainian and one HIT for English→Czech.

German→English

Rank	Ave.	System
1-3	90.3	GPT4-5shot
1-3	89.9	Human-refA
1-5	89.6	ONLINE-A
3-6	89.1	ONLINE-B
3-6	88.8	ONLINE-W
4-7	88.0	ONLINE-Y
6-8	87.7	ONLINE-G
8-9	86.5	GTCOM_DLUT
7-9	85.3	ONLINE-M
10-11	81.8	LanguageX
10-13	80.0	Lan-BridgeMT
11-14	79.6	NLLB_MBR_BLEU
12-14	78.8	AIRC
11-14	77.9	NLLB_Greedy

English→German

Rank	Ave.	System
1-5	89.0	GPT4-5shot
1-5	88.8	ONLINE-B
1-4	88.3	ONLINE-W
2-6	88.1	ONLINE-A
4-6	88.0	ONLINE-Y
1-6	87.7	Human-refA
7-8	86.7	ONLINE-M
7-8	85.5	ONLINE-G
9	84.0	Lan-BridgeMT
10	82.7	LanguageX
11-12	76.8	NLLB_MBR_BLEU
11-12	75.7	NLLB_Greedy
13	73.6	AIRC

English→Czech

Rank	Ave.	System
1	85.4	Human-refA
2	84.1	ONLINE-W
3-5	81.8	GPT4-5shot
3-4	80.4	CUNI-GA
5-8	80.3	ONLINE-A
5-8	79.4	CUNI-DocTransformer
4-7	78.8	ONLINE-B
8-14	78.6	NLLB_MBR_BLEU
6-11	78.4	GTCOM_DLUT
8-12	77.4	CUNI-Transformer
10-14	76.8	NLLB_Greedy
9-14	75.7	ONLINE-M
10-15	75.2	ONLINE-G
13-15	75.0	ONLINE-Y
8-15	75.0	Lan-BridgeMT
16	74.1	LanguageX

Czech→Ukrainian

Rank	Ave.	System
1-3	83.7	ONLINE-B
1-3	83.6	GPT4-5shot
1-3	83.2	Human-refA
4-8	82.8	ONLINE-W
4-8	82.4	CUNI-GA
4-8	81.8	CUNI-Transformer
4-8	81.3	GTCOM_DLUT
4-8	80.6	ONLINE-A
9-11	79.5	ONLINE-G
9-13	78.7	ONLINE-Y
9-13	78.7	MUNI-NLP
10-13	77.4	Lan-BridgeMT
10-13	76.9	NLLB_MBR_BLEU
14	76.7	NLLB_Greedy

Chinese→English

Rank	Ave.	System
1-2	82.9	Lan-BridgeMT
1-2	80.9	GPT4-5shot
3-8	80.3	Yishu
3-7	80.2	ONLINE-W
5-10	80.0	ONLINE-G
3-7	79.8	ONLINE-B
4-9	79.7	ONLINE-Y
3-8	79.1	HW-TSC
6-10	77.8	ONLINE-A
10-11	77.7	IOL_Research
8-11	77.2	LanguageX
12-13	76.9	ONLINE-M
13-16	76.2	NLLB_MBR_BLEU
12-15	76.1	Human-refA
14-16	74.0	NLLB_Greedy
13-16	72.6	ANVITA

English→Chinese

Rank	Ave.	System
1-5	82.2	Yishu
1-5	82.1	Human-refA
1-7	82.1	GPT4-5shot
3-8	82.0	Lan-BridgeMT
1-6	81.8	ONLINE-B
1-8	81.5	HW-TSC
4-8	81.4	ONLINE-W
5-8	80.2	ONLINE-Y
9-10	79.8	IOL_Research
9-10	79.7	ONLINE-A
11-13	78.6	LanguageX
11-13	78.2	ONLINE-M
11-13	77.1	ONLINE-G
14	64.5	ANVITA
15	64.3	NLLB_Greedy
16	57.2	NLLB_MBR_BLEU

Japanese→English

Rank	Ave.	System
1	81.3	GPT4-5shot
2-4	80.6	SKIM
3-8	80.4	Human-refA
3-8	79.5	ONLINE-Y
2-8	79.4	ONLINE-B
3-9	79.2	ONLINE-A
2-8	78.8	ONLINE-W
3-8	78.4	NAIST-NICT
8-9	76.9	GTCOM_DLUT
10-13	76.4	Lan-BridgeMT
10-13	75.8	ANVITA
10-13	74.8	ONLINE-G
10-13	74.6	LanguageX
14-15	72.9	ONLINE-M
14-15	72.4	KYB
16	68.9	AIRC
17-18	66.7	NLLB_MBR_BLEU
17-18	66.1	NLLB_Greedy

English→Japanese

Rank	Ave.	System
1-2	80.7	Human-refA
2-6	79.5	GPT4-5shot
1-5	78.8	ONLINE-B
2-6	78.6	ONLINE-Y
2-5	78.5	SKIM
4-6	78.4	ONLINE-W
7-10	76.6	LanguageX
7-10	76.2	ONLINE-A
7-10	76.1	NAIST-NICT
7-10	75.2	Lan-BridgeMT
11-12	73.1	ANVITA
11-12	72.6	ONLINE-M
13-15	70.8	KYB
13-15	69.6	AIRC
13-15	69.6	ONLINE-G
16	64.5	NLLB_Greedy
17	61.3	NLLB_MBR_BLEU

Table 6: Official results of WMT23 General Translation Task. Systems ordered by DA score; systems within a cluster are considered tied; lines indicate clusters according to Wilcoxon rank-sum test $p < 0.05$; rank ranges indicate the number of systems a system significantly underperforms or outperforms; grayed entry indicates resources that fall outside the constraints provided. All language pairs used document-level evaluation.

systems evenly across annotators. This approach aims to minimize the potential bias of a particularly stringent annotator disproportionately penalizing a single system. Ideally, every annotator would assess documents translated by all systems; however, this could introduce task repetitiveness concerns. For future considerations, employing calibration HITs (see Section 5.3) to normalize each annotator's behaviour could offer a promising solution.

All segment-level scores are averaged per system to compute the system-level scores. The clusters are computed using the Wilcoxon rank-sum test with $p < 0.05$. Rank ranges indicate the number of systems a particular system underperforms or outperforms: the top end of the rank range is $l + 1$ where l is the number of losses, while the bottom is $n - w$ where n is the total number of systems and w is the number of systems that the system in questions significantly wins against.

Tables with head-to-head comparisons between all systems are included in Appendix G.

At the time of preparation of the camera-ready version of the paper, we have not been able to collect the required number of high-quality assessments for language pairs run through Toloka AI that would meet WMT standards for human evaluation. In that regard, we decided not to publish official rankings based on manual evaluation for English↔{Hebrew, Russian, Ukrainian} until the conference, we are planning to address it later.

5.5 Comparison of human evaluation methods

In collaboration with the metrics shared task (Freitag et al., 2023), human annotation data for the Chinese→English and English→German direction was collected using two different approaches: the source-based DA+SQM approach, and the Multi-dimensional Quality Metrics (MQM) framework (Freitag et al., 2021). We present the rankings produced by the two approaches in Table 7.

Upon examining the system rankings and individual clusters produced by both techniques, it is evident that DA+SQM produces fewer clusters. This suggests that it might not be sufficiently robust to differentiate smaller system differences, whereas MQM creates more detailed clusters. One potential explanation is that DA+SQM, constrained by budgetary restrictions, might be under-powered. As highlighted by Wei et al. (2022), the 1500 segments we gather per system might not suffice to segregate systems in a more detailed manner.

Conversely, the largest difference in the evaluation techniques is the cost. While MQM manages to establish more refined clusters, its deployment is significantly more costly and complex, especially when training professionals. An interesting question would be determining the number of MQM labels that could be procured within the budget allocated for DA+SQM.

It is also important to note that the set of data over which each of these rankings was produced may have differed slightly due to the sampling (e.g., the distribution over topic domains or the amount of coverage of the full test set), making it difficult to determine whether these differences in rankings represent differences due to data or due to different annotation methods.

6 Test Suites

As can be seen in the general MT task, the improvement of translation quality has made it difficult to discriminate MT output from human translation with the current evaluation methods. Nevertheless, there are still cases where MT has difficulties, delivering outputs which despite seeming fluent and being surrounded by other seemingly perfect translations, entail serious flaws. In general evaluation methods, such flaws can get "hidden in the average" or simply get missed altogether. In an effort to shed light to these cases, evaluation via test suites is embedded in the shared task.

6.1 Setup of the sub-task

Test suites are custom extensions to standard test sets, constructed so that they can focus on particular aspects of the MT output. Here, the evaluation of the MT outputs takes place in a decentralized manner as a part of a sub-task, where test suite providers were invited to submit their customized test sets, following the setting introduced at the Third Conference on Machine Translation (Bojar et al., 2018).

Every test suite provider submitted a source-side test set, which the shared task organizers appended to the standard test sets of the shared task. The corresponding outputs from the MT systems of the shared task were returned to the test suite providers, who were responsible for running the evaluation, based on their own custom evaluation methods. The results of each test suite evaluation, together with the relevant analysis, appear in separate description papers.

Rank	Ave. ↑	System (En-De)
1-5	89.0	GPT4-5shot
1-5	88.8	ONLINE-B
1-4	88.3	ONLINE-W
2-6	88.1	ONLINE-A
4-6	88.0	ONLINE-Y
1-6	87.7	Human-refA
7-8	86.7	ONLINE-M
7-8	85.5	ONLINE-G
9	84.0	Lan-BridgeMT
10	82.7	LanguageX
11-12	76.8	NLLB_MBR_BLEU
11-12	75.7	NLLB_Greedy
13	73.6	AIRC

System (En-De)	MQM ↓
refA	2.96
GPT4-5shot	3.72
ONLINE-W	3.95
ONLINE-B	4.71
ONLINE-Y	5.64
ONLINE-A	5.67
ONLINE-G	6.57
ONLINE-M	6.94
Lan-BridgeMT	8.67
LanguageX	9.25
NLLB_Greedy	9.54
NLLB_MBR_BLEU	10.79
AIRC	14.23

Rank	Ave. ↑	System (Zh-En)
1-2	82.9	Lan-BridgeMT
1-2	80.9	GPT4-5shot
3-8	80.3	Yishu
3-7	80.2	ONLINE-W
5-10	80.0	ONLINE-G
3-7	79.8	ONLINE-B
4-9	79.7	ONLINE-Y
3-8	79.1	HW-TSC
6-10	77.8	ONLINE-A
10-11	77.7	IOL_Research
8-11	77.2	LanguageX
12-13	76.9	ONLINE-M
13-16	76.2	NLLB_MBR_BLEU
12-15	76.1	Human-refA
14-16	74.0	NLLB_Greedy
13-16	72.6	ANVITA

System (Zh-En)	MQM ↓
Lan-BridgeMT	2.10
GPT4-5shot	2.31
Yishu	3.23
ONLINE-B	3.39
HW-TSC	3.40
ONLINE-A	3.79
ONLINE-Y	3.79
ONLINE-G	3.86
ONLINE-W	4.06
LanguageX	4.23
IOL_Research	4.59
refA	4.83
ONLINE-M	5.43
ANVITA	6.08
NLLB_MBR_BLEU	6.36
NLLB_Greedy	6.57

Table 7: Comparison of system clustering as done by DA+SQM and MQM technique. Top two tables are for English to German, while bottom two are for Chinese to German.

6.2 Submissions

The test suite sub-task received 5 submissions with 6 test suites, whose overview can be seen in Table 8. The descriptions of each submission and their main findings are given below.

DFKI (Manakhimova et al., 2023) test suite offers a fine-grained linguistically motivated analysis of the shared task MT outputs, based on more than 11,500 manually devised test items, which cover up to 110 phenomena in 14 categories per language direction. Extending their previous test suite efforts (e.g. Avramidis et al., 2018; Macketanz et al., 2022), the submission of this year includes an updated test set featuring new linguistic phenomena and focuses additionally on the participating LLMs. The evaluation spans German→English, English→German, and English→Russian language directions.

Some of the phenomena with the lowest accuracies for German→English are *idioms* and *resultative predicates*. For English→German, these include *mediopassive voice*, and *noun formation(er)*. As for English→Russian, these include *idioms* and *semantic roles*. GPT4 performs equally or comparably to the best systems in German→English and English→German but falls in the second significance cluster for English→Russian.

HW-TSC (Chen et al., 2023) propose a systematic approach to select test sentences with high-level of difficulty from the Wiki Corpus. The strategy considers the difficulty level of a sentence from four dimensions: *word difficulty, length difficulty, grammar difficulty* and *model learning difficulty*. They open-source two Multifaceted Challenge Sets for Chinese→English and English→Chinese, each of them containing 2,000 sentences. Then, they use these challenge sets to test the shared task systems, presenting results by three automatic metrics.

The resulting system ranks are quite different from the official results. The authors point out that systems that perform well on average test sets may not perform as well on sets with high difficulty. If the ranking difference is caused by domain issues, the top-ranked systems on the official test sets may not be so general. GPT4 is ranked in the first two positions in Chinese→English but its rank in

Test suite	Directions	Phenomena	#Sentences	Citation	Link
DFKI	de–en, en–de, en–ru	110 linguistic phenomena	11,517	Manakhimova et al. (2023)	DFKI-NLP
HW-TSC	zh–en, en–zh	4 difficulty dimensions	4,000	Chen et al. (2023)	HwTsc
IIIT HYD	en–de	5 domains, 5 writing styles	2,268	Mukherjee and Shrivastava (2023)	wmt23
INES	en–de	Inclusive language forms	162	Savoldi et al. (2023)	fbk.eu
MuST-SHE	en–de	Binary gender bias	200	Savoldi et al. (2023)	fbk.eu
RoCS-MT	en–de, en–cs, en–uk, en–ru	Non-standard user-generated content	1,922	Bawden and Sagot (2023)	RoCS-MT

Table 8: Overview of the participating test suites.

English→Chinese is much lower (ranks 4-9).

IIIT HYD (Mukherjee and Shrivastava, 2023) This test suite covers five specific domains (entertainment, environment, health, science, legal) and spans five distinct writing styles (descriptive, judgments, narrative, reporting, technical-writing) for English–German. The authors conduct their analysis through a combination of au- tomated assessments and manual evaluations.

Based on their evaluation, it is evident that both ONLINE-B and ONLINE-Y consistently surpassed other MT systems in performance across a diverse array of writing styles and domains. When focusing on GPT4, whereas it performs comparably to the best systems for most domains and writing styles, it gives considerably worse results when applied to the legal domain, and the writing style of judgments.

MuST-SHE^{WMT23} and INES (Savoldi et al., 2023) By focusing on the en-de and de-en language pairs, the authors rely on these newly created test suites to investigate systems' ability to translate feminine and masculine gender and produce gender-inclusive translations. Furthermore, they discuss metrics associated with the test suites and validate them by means of human evaluations.

The results indicate that systems achieve reasonable and comparable performance in correctly translating both feminine and masculine gender forms for naturalistic gender phenomena. Instead, the generation of inclusive language forms in translation emerges as a challenging task for all the evaluated MT models, indicating room for future improvements and research on the topic.

Concerning GPT 4, it is noticeable that its overall accuracy is 2% worse than the best MT system, whereas it achieves a relatively low accuracy with regard to the feminine gender, when evaluating whether the first-person singular references to the speaker are translated according to the speaker's linguistic expression of gender.

RoCS-MT (Bawden and Sagot, 2023) The RoCS-MT Challenge Set is designed to test MT systems' robustness to user-generated content (UGC) displaying non-standard characteristics, such as spelling errors, devowelling, acronymisation, etc. It is composed of non-standard English comments from Reddit, manually normalised and professionally translated into four of the WMT 2023 target languages, German, Czech, Ukrainian and Russian, and also French.

Through automatic and manual analysis of system outputs, we find that many of the phenomena remain challenging for most systems, but to varying degrees depending on the phenomenon, the particular instance (notably how frequent the non-standard word is) and the system, especially with respect to the quantity of training data. For example, non-standard instances of words (e.g. through devowelling or through phonetically inspired spelling) are often either omitted in the translation or copied unchanged. When non-standard words are translated, it is often in their standard form, but with some exceptions, for example capitalisation is sometimes preserved. However, there is often inconsistency within a same system's outputs.

GPT4-5shot has a clear lead over all other systems, correctly translating even some of the most challenging examples. It sometimes (although inconsistently) reproduces non-standardness in its outputs, but also does not always remain entirely faithful to the source sentence. However, aside the huge disparity in the amount of training data compared to other systems, notably the constrained ones, the lack of access to its training data is a serious obstacle to any meaningful scientific comparison; we cannot know which phenomena were seen during training and how frequently, and more

crucially, we cannot verify whether RoCS-MT sentences were seen during training.

7 Conclusions

The General Machine Translation Task at WMT 2023 covered 14 translation pairs, where the only non-English language pair was Czech→Ukrainian. Source based DA+SQM was the main human golden truth. The evaluation included 72 primary submissions from 17 participants, 6 online systems and 3 additional contrastive systems including GPT4. It was performed by 155 human (semi-)professional annotators, who contributed more than 175,000 judgments altogether. For most language pairs (apart from English→Czech), MT systems produce outputs that cannot be identified as being worse than the manually produced references translations in a statistically significant way, using our current evaluation methods.

It is apparent that this year, the amount of unconstrained submissions are lower thank in past years (27 submissions by 11 participants). Additionally, for some language pairs there are only few submissions by participants, and therefore they are dominated by many online systems, of whom we have no technical descriptions. We are therefore considering ways to encourage participation in the future, whereas redefining the constrained setting may be needed.

It is the first time that Large Language Models (LLMs) are included in the Shared Task as translation systems. Although the technology is very apparent in NLP research, we received only one submission using LLM methods (Lan-BridgeMT), whereas one dominant commercial LLM (GPT4) was included via our own efforts. GPT4 was in the first significance cluster for all systems translating towards English, but fell in the second significance cluster (rank 3-5) for English→Czech, whereas a similar sign was given by one of the test suites for English→Russian (rank 3; Manakhimova et al., 2023). Additionally, test suites providers noted that GPT4 outputs are not always faithful to the source sentence (Bawden and Sagot, 2023) and that they have some issues with speaker gender translation (Savoldi et al., 2023) and specific domains (Mukherjee and Shrivastava, 2023, e.g. legal;). Due to the closed-source nature of commercial tools, it is hard to know the exact reasons for these findings, although they confirm previous observations that GPT models have difficulties with under-represented languages (Hendy et al., 2023). We believe that a more transparent comparison including open source LLMs should be sought for the future.

8 Limitations

We investigated a research question of testing general capabilities of MT systems. However, we have simplified this approach. Firstly, we only used four domains that are not specialized. Secondly, we used only cleaner sentences, avoiding noisy in the source sentences.

Although we accept human judgement as a gold standard, giving us more reliable signal than automatic metrics, we should mention that human annotations are noisy (Wei and Jia, 2021) and their performance is affected by quality of other evaluated systems (Mathur et al., 2020).

Different annotators are using different ranking strategy which may have an effect on the system ranking as we are using raw scores.

9 Ethical Consideration

Several of the domains contained texts that included personal data, for example the speech data (See Section 2.4 for more details). Entities were replaced by anonymisation tags (e.g. #NAME#, #EMAIL#) to preserve the anonymity of the users behind the content.

The sentences in Ukrainian datasets were collected with users' opt-in consent, and any personal data related to people other than well-known people was pseudonymized (using random first names and surnames). Sentences where such pseudonymization would not be enough to preserve reasonable anonymity of the users (e.g. describing events uniquely identifying the persons involved) were not included in the test set.

As described in Section 2.2 and in the linguistic brief (Appendix Section B), inappropriate, controversial and/or explicit content was filtered out prior to translation, particularly keeping in mind the translators and not exposing them to such content or obliging them to translate it. A few sentences containing explicit content managed to escape the filter, and we removed these sentences from the test sets without translation.

Human evaluation using Appraise for collecting human judgements was fully anonymous. Automatically generated accounts associated with annotation tasks with single-sign-on URLs were dis-

tributed randomly among pools of annotators and did not allow for storing personal information. For language pairs for which we used calibration HITs, we received lists of tasks completed by an individual anonymous annotator. Annotators have been well paid in respect to their countries.

Acknowledgments

This task would not have been possible without the sponsorship of monolingual data, test sets translation and evaluation from our partners. Namely Microsoft, Charles University, Toloka AI, Google, NTT Resonant, Dubformer, and Centific.

Additionally, we would like to thank Rebecca Knowles, Sergio Bruccoleri, Mariia Anisimova and many others who provided help and recommendations.

Barry Haddow's participation was funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10039436 – UTTER].

Rachel Bawden's participation was funded by her chair position in the PRAIRIE institute funded by the French national agency ANR as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001 and by her Emergence project, DadaNMT, funded by Sorbonne Université.

Maja Popović's participation was funded by the ADAPT SFI Centre for Digital Media Technology, funded by Science Foundation Ireland through the SFI Research Centres Programme and co-funded under the European Regional Development Fund (ERDF) through Grant 13/RC/2106.

Martin Popel's participation was funded by GAČR EXPRO grant LUSyD (GX20-16819X).

Eleftherios Avramidis's participation was funded by the German Research Foundation (DFG) through the project TextQ (grant num. MO 1038/31-1, 436813723), and by the German Federal Ministry of Education and Research (BMBF) through the project SocialWear (grant num. 01IW2000).

This work has been using data and tools provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Eleftherios Avramidis, Vivien Macketanz, Arle Lommel, and Hans Uszkoreit. 2018. Fine-grained evaluation of quality estimation for machine translation based on a linguistically motivated test suite. In *Proceedings of the AMTA 2018 Workshop on Translation Quality Estimation and Automatic Post-Editing*, pages 243–248, Boston, MA. Association for Machine Translation in the Americas.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Rachel Bawden and Benoît Sagot. 2023. RoCS-MT: Robustness Challenge Set for Machine Translation. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala, Sweden. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland. Association for Computational Linguistics.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy. European Association for Machine Translation.

Xiaoyu Chen, Daimeng Wei, Zhanglin Wu, Ting Zhu, Hengchao Shang, Zongyao Li, Jiaxin GUO, Ning Xie, Lizhi Lei, Hao Yang, and Yanfei Jiang. 2023. Multifaceted Challenge Set for Evaluating Machine Translation Performance. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jan Christian Blaise Cruz. 2023. Samsung R&D Institute Philippines at WMT 2023. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Hiroyuki Deguchi, Kenji Imamura, Yuto Nishida, Yusuke Sakai, Justin Vasselli, and Taro Watanabe. 2023. NAIST-NICT WMT'23 general mt task submission. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Christian Federmann. 2018. Appraise evaluation framework for machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. APE at scale and its implications on MT evaluation biases. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Nitika Mathur, Chi kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frédéric Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 Metrics Shared Task: Metrics might be Guilty but References are not Innocent. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Tirthankar Ghosal, Marie Hledíková, Muskaan Singh, Anna Nedoluzhko, and Ondřej Bojar. 2022. The second automatic minuting (AutoMin) challenge: Generating and evaluating minutes from multi-party meetings. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pages 1–11, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Thamme Gowda, Zhao Zhang, Chris Mattmann, and Jonathan May. 2021. Many-to-English machine translation tools, data, and pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 306–316, Online. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Josef Jon, Martin Popel, and Ondřej Bojar. 2023. CUNI-GA submission at WMT23. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4563–4568, Online. Association for Computational Linguistics.

Ondrej Klejch, Eleftherios Avramidis, Aljoscha Burchardt, and Martin Popel. 2015. MT-ComparEval: Graphical evaluation interface for machine translation development. *Prague Bull. Math. Linguistics*, 104:63–74.

Rebecca Knowles. 2021. On the stability of system rankings at WMT. In *Proceedings of the Sixth Conference on Machine Translation*, pages 464–477, Online. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Tom Kocmi, Martin Popel, and Ondrej Bojar. 2020. Announcing CzEng 2.0 Parallel Corpus with over 2 Gigawords. *CoRR*, abs/2007.03006.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Philipp Koehn and Christof Monz, editors. 2006. *Proceedings on the Workshop on Statistical Machine Translation*. Association for Computational Linguistics, New York City.

Keito Kudo, Takumi Ito, Makoto Morishita, and Jun Suzuki. 2023. SKIM at WMT 2023 general translation task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Ben LI, Yoko Matsuzaki, and Shivam Kalkar. 2023. KYB general machine translation systems for WMT23. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A Set of Recommendations for Assessing Human–Machine Parity in Language Translation. *Journal of Artificial Intelligence Research (JAIR)*, 67.

Xuezhe Ma, Chunting Zhou, Xiang Kong, Junxian He, Liangke Gui, Graham Neubig, Jonathan May, and Zettlemoyer Luke. 2022. Mega: Moving average equipped gated attention. *arXiv preprint arXiv:2209.10655*.

Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, He Wang, Renlong Ai, Shushen Manakhimova, Ursula Strohriegel, Sebastian Möller, and Hans Uszkoreit. 2022. A linguistically motivated test suite to semi-automatically evaluate German–English machine translation output. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 936–947, Marseille, France. European Language Resources Association.

Shushen Manakhimova, Eleftherios Avramidis, Vivien Macketanz, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2023. Linguistically motivated Evaluation of the 2023 State-of-the-art Machine Translation: Can ChatGPT Outperform NMT? In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Luo Min, yixin tan, and Qiulin Chen. 2023. Yishu: Yishu at WMT2023 translation task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Alexander Molchanov and Vladislav Kovalenko. 2023. PROMT systems for WMT23 shared general translation task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2020. JParaCrawl: A large scale web-based English-Japanese parallel corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3603–3609, Marseille, France. European Language Resources Association.

Ananya Mukherjee and Manish Shrivastava. 2023. IIIT HYD's Submission for WMT23 Test-suite task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Mathias Müller, Malihe Alikhani, Eleftherios Avramidis, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Sarah Ebling, Cristina España-bonet, Anne Göhring, Roman Grundkiewicz, Mert Inan, Zifan Jiang, Oscar Koller, Amit Moryossef, Annette Rios, Dimitar Shterionov, Sandra Sidler-miserez, Katja Tissi, and Davy Van Landuyt. 2023. Findings of the second WMT shared task on sign language translation (WMT-SLT23). In *Proceedings of the Eight Conference on Machine Translation (WMT)*, Singapore. Association for Computational Linguistics.

Anna Nedoluzhko, Muskaan Singh, Marie Hledíková, Tirthankar Ghosal, and Ondřej Bojar. 2022. ELITR minuting corpus: A novel dataset for automatic minuting from multi-party meetings in English and Czech. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3174–3182, Marseille, France. European Language Resources Association.

Graham Neubig. 2011. The Kyoto free translation task. http://www.phontron.com/kftt.

Mariana Neves, Antonio Jimeno Yepes, Aurélie Névéol, Rachel Bawden, Giorgio Maria Di Nunzio, Roland Roller, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Lana Yeganova, Dina Wiemann, and Cristian Grozea. 2023. Findings of the WMT 2023 Biomedical Translation Shared Task: Evaluation of ChatGPT 3.5 as a Comparison System. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv preprint arXiv:2207.04672*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Stephan Peitz, Sarthak Garg, Udhay Nallasamy, and Matthias Paulik. 2019. Cross+Self-Attention for Transformer Models. `https://github.com/pytorch/fairseq/files/3561282/paper.pdf`.

Martin Popel. 2020. CUNI English-Czech and English-Polish systems in WMT20: Robust document-level training. In *Proceedings of the Fifth Conference on Machine Translation*, pages 269–273, Online. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Reid Pryzant, Youngjoo Chung, Dan Jurafsky, and Denny Britz. 2018. JESC: Japanese-English subtitle corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Matīss Rikters. 2018. Impact of Corpora Quality on Neural Machine Translation. In *In Proceedings of the 8th Conference Human Language Technologies - The Baltic Perspective (Baltic HLT 2018)*, Tartu, Estonia. IOS Press.

Matīss Rikters and Makoto Miwa. 2023. AIST AIRC submissions to the WMT23 shared task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Roberts Rozis and Raivis Skadiņš. 2017. Tilde MODEL - multilingual open data for EU languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 263–265, Gothenburg, Sweden. Association for Computational Linguistics.

Pavel Rychlý and Yuliia Teslia. 2023. MUNI-NLP submission for Czech-Ukrainian translation task at WMT23. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Matteo Negri, and Luisa Bentivogli. 2023. Test Suites Task: Evaluation of Gender Fairness in MT with MuST-SHE and INES. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Roman Sudarikov, Martin Popel, Ondřej Bojar, Aljoscha Burchardt, and Ondřej Klejch. 2016. Using MT-ComparEval. In *Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem*, pages 76–82.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Christopher Lemmer Webber, Jessica Tallon, Erin Shepherd, Amy Guy, and Evan Prodromou. 2018. ActivityPub, W3C Recommendation. Technical report, W3C.

Johnny Wei and Robin Jia. 2021. The statistical advantage of automatic NLG metrics at the system level. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6840–6854, Online. Association for Computational Linguistics.

Johnny Wei, Tom Kocmi, and Christian Federmann. 2022. Searching for a higher power in the human evaluation of MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 129–139, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Di Wu and Christof Monz. 2023. Beyond shared vocabulary: Increasing representational word similarities across languages for multilingual machine translation. *arXiv preprint arXiv:2305.14189*.

Di Wu, Shaomu Tan, David Stap, Ali Araabi, and Christof Monz. 2023a. UvA-MT's participation in the WMT 2023 general translation shared task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Yangjian Wu and Gang Hu. 2023. Exploring prompt engineering with GPT language models for document-level machine translation: Insights and findings. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Zhanglin Wu, Daimeng Wei, Zongyao Li, Zhengzhe YU, Shaojun Li, Xiaoyu Chen, Hengchao Shang, Jiaxin GUO, Yuhao Xie, Lizhi Lei, Hao Yang, and Yanfei Jiang. 2023b. Treating general mt shared task as a multi-domain adaptation problem: Hw-tsc's submission to the WMT23 general mt shared task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Yiming Yan, Tao Wang, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Mingxuan Wang. 2023. BLEURT has universal translations: An analysis of automatic metrics by minimum risk training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5428–5443, Toronto, Canada. Association for Computational Linguistics.

Hui Zeng. 2023. Achieving state-of-the-art multilingual translation model with minimal data and parameters. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Wenbo Zhang. 2023. IOL Research machine translation systems for WMT23 general machine translation shared task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

Hao Zong. 2023. GTCOM neural machine translation systems for WMT23. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

A Statistics of training data

This section describes statistics of the training corpora.

Dataset ID	Segs	Tokens		Chars	
eng-ces		**eng**	**ces**	**eng**	**ces**
Facebook-wikimatrix-1-ces-eng	2.09M	33.56M	29.66M	206.82M	216.62M
ParaCrawl-paracrawl-9-eng-ces	50.63M	692.12M	626.34M	4.33B	4.68B
Statmt-commoncrawl_wmt13-1-ces-eng	161.84k	3.35M	2.93M	20.66M	20.75M
Statmt-europarl-10-ces-eng	644.43k	15.63M	13.00M	94.31M	98.14M
Statmt-news_commentary-16-ces-eng	253.27k	5.46M	4.96M	34.58M	37.97M
Statmt-wikititles-3-ces-eng	410.94k	1.03M	965.62k	7.47M	7.57M
Tilde-ecb-2017-ces-eng	3.10k	52.12k	45.21k	327.57k	339.24k
Tilde-eesc-2017-ces-eng	1.33M	28.78M	25.63M	188.53M	205.14M
Tilde-ema-2016-ces-eng	495.23k	7.64M	7.28M	50.31M	57.01M
Tilde-rapid-2019-ces-eng	263.29k	5.79M	5.30M	37.36M	41.26M
(Total)	56.29M	793.41M	716.10M	4.97B	5.36B
eng-deu		**eng**	**deu**	**eng**	**deu**
Facebook-wikimatrix-1-deu-eng	6.23M	100.50M	96.95M	623.66M	701.23M
ParaCrawl-paracrawl-9-eng-deu	278.31M	4.27B	3.99B	26.37B	29.46B
Statmt-commoncrawl_wmt13-1-deu-eng	2.40M	51.40M	47.05M	314.18M	340.51M
Statmt-europarl-10-deu-eng	1.82M	45.51M	42.41M	272.94M	312.14M
Statmt-news_commentary-16-deu-eng	388.48k	8.55M	8.77M	54.40M	65.94M
Statmt-wikititles-3-deu-eng	1.47M	3.61M	3.08M	26.48M	25.50M
Tilde-airbaltic-1-deu-eng	0.84k	17.60k	15.08k	104.34k	105.52k
Tilde-czechtourism-1-deu-eng	6.76k	128.29k	114.44k	769.04k	829.41k
Tilde-ecb-2017-deu-eng	4.15k	85.52k	74.81k	545.51k	582.63k
Tilde-eesc-2017-deu-eng	2.86M	61.47M	58.28M	400.37M	469.94M
Tilde-ema-2016-deu-eng	347.63k	5.09M	5.01M	33.48M	39.43M
Tilde-rapid-2016-deu-eng	1.03M	20.65M	19.85M	134.26M	158.13M
Tilde-rapid-2019-deu-eng	939.81k	19.90M	19.30M	129.03M	153.08M
(Total)	295.81M	4.59B	4.29B	28.36B	31.73B
eng-heb		**eng**	**heb**	**eng**	**heb**
ELRC-wikipedia_health-1-eng-heb	3.16k	69.71k	54.76k	442.38k	583.87k
Facebook-wikimatrix-1-eng-heb	2.04M	35.83M	28.96M	218.77M	300.61M
Neulab-tedtalks_train-1-eng-heb	211.82k	4.45M	3.44M	22.36M	29.00M
OPUS-bible_uedin-v1-eng-heb	62.20k	1.55M	830.23k	8.16M	7.46M
OPUS-ccmatrix-v1-eng-heb	25.23M	313.87M	249.49M	1.81B	2.45B
OPUS-elrc_2922-v1-eng-heb	3.16k	69.73k	54.77k	442.40k	583.54k
OPUS-elrc_3065_wikipedia_health-v1-eng-heb	3.16k	69.71k	54.76k	442.31k	583.51k
OPUS-elrc_wikipedia_health-v1-eng-heb	3.16k	69.71k	54.76k	442.31k	583.51k
OPUS-globalvoices-v2018q4-eng-heb	1.03k	20.31k	15.03k	122.39k	158.63k
OPUS-gnome-v1-eng-heb	0.15k	0.42k	0.40k	2.89k	3.96k
OPUS-kde4-v2-eng-heb	79.32k	338.22k	347.35k	2.09M	3.13M
OPUS-multiccaligned-v1-eng-heb	5.33M	60.55M	52.81M	380.74M	518.33M
OPUS-opensubtitles-v2018-eng-heb	29.89M	195.98M	154.25M	1.03B	1.40B
OPUS-php-v1-eng-heb	27.82k	83.46k	93.03k	498.72k	789.34k
OPUS-qed-v2.0a-eng-heb	464.35k	6.37M	4.48M	34.70M	42.34M
OPUS-tatoeba-v20220303-eng-heb	164.20k	1.02M	806.38k	5.41M	7.37M
OPUS-tatoeba-v2-eng-heb	54.36k	357.09k	277.32k	1.87M	2.56M
OPUS-ubuntu-v14.10-eng-heb	1.44k	6.13k	5.78k	38.78k	54.69k
OPUS-wikimedia-v20210402-eng-heb	226.83k	8.51M	7.56M	57.58M	78.26M
OPUS-wikipedia-v1.0-eng-heb	139.85k	2.69M	2.27M	16.45M	22.43M
OPUS-xlent-v1.1-eng-heb	3.19M	9.61M	7.93M	60.53M	73.11M
Statmt-ccaligned-1-eng-heb_IL	5.33M	60.55M	52.81M	380.76M	518.34M
(Total)	72.46M	702.05M	566.59M	4.04B	5.45B

Table 9: Statistics for parallel training set provided for General/News Translation Task. Suffixes, k, M, and B, are short for thousands, millions, and billions, respectively. Dataset ID is the unique identifier created by MTData, example `mtdata echo <dataset_id>`.

Dataset ID	Segs	Tokens		Chars	
eng-jpn		**eng**		**eng**	**jpn**
Facebook-wikimatrix-1-eng-jpn	3.90M	61.63M		379.09M	454.97M
KECL-paracrawl-3-eng-jpn	25.74M	599.02M		3.69B	4.58B
Phontron-kftt_train-1-eng-jpn	440.29k	9.74M		59.91M	49.08M
StanfordNLP-jesc_train-1-eng-jpn	2.80M	19.34M		104.00M	119.62M
Statmt-news_commentary-16-eng-jpn	1.84k	39.50k		247.70k	310.56k
Statmt-ted-wmt20-eng-jpn	241.74k	4.03M		23.02M	27.32M
Statmt-wikititles-3-jpn-eng	757.04k	1.94M		13.96M	18.67M
(Total)	33.88M	695.74M		4.27B	5.25B
eng-rus		**eng**	**rus**	**eng**	**rus**
Facebook-wikimatrix-1-eng-rus	5.20M	86.79M	76.48M	537.73M	965.44M
OPUS-unpc-v1.0-eng-rus	25.17M	563.82M	520.71M	3.70B	7.31B
ParaCrawl-paracrawl-1_bonus-eng-rus	5.38M	101.31M	80.41M	632.54M	1.06B
Statmt-backtrans_enru-wmt20-eng-rus	36.77M	736.20M	670.93M	4.31B	7.73B
Statmt-commoncrawl_wmt13-1-rus-eng	878.39k	18.77M	17.40M	116.16M	214.59M
Statmt-news_commentary-16-eng-rus	331.51k	7.67M	7.13M	48.79M	97.41M
Statmt-wikititles-3-rus-eng	1.19M	3.13M	2.88M	22.80M	39.34M
Statmt-yandex-wmt22-eng-rus	1.00M	21.25M	18.68M	130.99M	250.76M
Tilde-airbaltic-1-eng-rus	1.09k	23.98k	18.79k	142.52k	252.73k
Tilde-czechtourism-1-eng-rus	7.33k	140.09k	110.10k	838.09k	1.50M
Tilde-worldbank-1-eng-rus	25.85k	588.58k	573.93k	3.85M	8.21M
(Total)	75.96M	1.54B	1.40B	9.50B	17.67B
eng-ukr		**eng**	**ukr**	**eng**	**ukr**
ELRC-acts_ukrainian-1-eng-ukr	129.94k	3.04M	2.60M	19.55M	35.69M
Facebook-wikimatrix-1-eng-ukr	2.58M	41.55M	35.59M	257.56M	447.33M
ParaCrawl-paracrawl-1_bonus-eng-ukr	13.35M	505.83M	487.47M	3.28B	6.04B
Tilde-worldbank-1-eng-ukr	1.63k	36.07k	34.18k	237.96k	477.91k
(Total)	16.06M	550.46M	525.68M	3.55B	6.52B
eng-zho		**eng**		**eng**	**zho**
Facebook-wikimatrix-1-eng-zho	2.60M	49.87M		311.07M	277.84M
OPUS-unpc-v1.0-eng-zho	17.45M	417.25M		2.75B	2.14B
ParaCrawl-paracrawl-1_bonus-eng-zho	14.17M	217.60M		1.34B	1.18B
Statmt-backtrans_enzh-wmt20-eng-zho	19.76M	364.22M		2.16B	1.96B
Statmt-news_commentary-16-eng-zho	313.67k	6.92M		44.14M	38.83M
Statmt-wikititles-3-zho-eng	921.96k	2.37M		17.82M	16.28M
(Total)	55.22M	1.06B		6.62B	5.61B
ces-ukr		**ces**	**ukr**	**ces**	**ukr**
ELRC-acts_ukrainian-1-ces-ukr	130.00k	2.48M	2.56M	19.61M	35.26M
Facebook-wikimatrix-1-ces-ukr	848.96k	10.43M	10.07M	75.97M	127.31M
OPUS-bible_uedin-v1-ces-ukr	7.95k	140.03k	132.06k	904.31k	1.33M
OPUS-ccmatrix-v1-ces-ukr	3.99M	45.13M	45.10M	330.68M	566.27M
OPUS-elrc_5179_acts_ukrainian-v1-ces-ukr	130.00k	2.48M	2.56M	19.61M	35.26M
OPUS-elrc_wikipedia_health-v1-ces-ukr	0.19k	3.23k	3.18k	24.27k	41.63k
OPUS-eubookshop-v2-ces-ukr	1.51k	23.71k	19.15k	187.30k	275.14k
OPUS-gnome-v1-ces-ukr	0.15k	0.42k	0.41k	3.53k	5.82k
OPUS-kde4-v2-ces-ukr	133.67k	593.82k	677.35k	4.45M	7.97M
OPUS-multiccaligned-v1.1-ces-ukr	1.61M	19.75M	19.77M	146.44M	244.36M
OPUS-multiparacrawl-v9b-ces-ukr	2.20M	25.62M	25.55M	188.08M	325.50M
OPUS-opensubtitles-v2018-ces-ukr	730.80k	3.88M	3.90M	24.20M	40.62M
OPUS-qed-v2.0a-ces-ukr	161.02k	2.02M	2.04M	13.44M	22.80M
OPUS-tatoeba-v20220303-ces-ukr	2.93k	10.85k	11.40k	68.70k	118.67k
OPUS-ted2020-v1-ces-ukr	114.23k	1.57M	1.56M	10.70M	17.93M
OPUS-ubuntu-v14.10-ces-ukr	0.23k	1.67k	1.76k	13.02k	20.86k
OPUS-wikimedia-v20210402-ces-ukr	1.96k	39.18k	34.91k	285.74k	414.20k
OPUS-xlent-v1.1-ces-ukr	695.41k	1.78M	1.58M	12.92M	18.30M
(Total)	10.76M	115.95M	115.57M	847.58M	1.44B

Table 10: Statistics for parallel training set provided for General/News Translation Task. Suffixes, k, M, and B, are short for thousands, millions, and billions, respectively.

B Preprocessing cleanup brief for linguists

Human check briefing

In this task, we wish to check the data to remove all inappropriate content, remove repetitive content, or correct minor problems with the text.

The data is automatically broken down into individual sentences, which may contain wrong sentence splitting that needs to be fixed. Each paragraph is separated by empty lines. Keep the document-separators intact.

We ask you to read each document and either:

- Delete document completely if it contains any of following issues. Be on the save side, rather remove documents where you are uncertain
    - Remove documents written in different language (natural code-switching is fine)
    - Remove inappropriate content (such as sexually explicit, vulgar, or otherwise inappropriate)
    - Remove controversial content (propagandist, controversial political topics, etc.)
    - Remove content that is too noisy or doesn't resemble natural text (such as documents badly formatted, hard to understand, containing unusual language, lists of numbers/data, or other structured data generated automatically)
- Keep document while checking
    - Fix sentence-breaking, each line must be one sentence (do not reformulate, simply remove or add end of lines on a proper place).
    - Remove or move fragments of sentences to previous or following sentence (for example emoticons, one or few words sentences)
    - Fix minor issues and keep it (do not spent too much time on fixing it).
        - It is fine to keep some errors or problems
        - Remove boilerplates (segments that break the document, for example ads, page numbers, signatures, artefacts, …)
    - If a given document has more than around 30 sentences, consider splitting it by adding an empty line on a meaningful place splitting it into paragraphs

This task shouldn’t take much longer than reading through documents.

C Translator Brief for General MT

Translator Brief

In this project we wish to translate online news articles for use in evaluation of Machine Translation (MT). The translations produced by you will be compared against the translations produced by a variety of different MT systems. They will be released to the research community to provide a benchmark, or "gold-standard" measure for translation quality. The translation therefore needs to be a high-quality rendering of the source text into the target language, as if it was news written directly in the target language. However, there are some constraints imposed by the intended usage:

- All translations should be "from scratch", without post-editing from MT. Using post-editing would bias the evaluation, so we need to avoid it. We can detect post-editing so will reject translations that are post-edited.
- Translation should preserve the sentence boundaries. The source texts are provided with exactly one sentence per line, and the translations should be the same, one sentence per line. Blank lines should be preserved in the translation.
- Translators should avoid inserting parenthetical explanations into the translated text and obviously avoid losing any pieces of information from the source text. We will check a sample of the translations for quality, and we will check the entire set for evidence of post-editing.
- Please do not translate the anonymization tags (e.g. #NAME#), but use the same form as in the source text. These tags are used to de-identify names and various other sensitive data. In other words, translation must contain given tag #NAME# on a position where it would naturally be placed before anonymization.
- If the original data contain errors, typos, or other problems, do not try to fix them (or introduce them in the translation), instead try to prepare correct translation as if the error wouldn't be in the source.

The source files will be delivered as text files (sometimes known as "notepad" files), with one sentence per line. We need the translations to be returned in the same format. If you prefer to receive the text in a different format, then please let us know as we may be able to accommodate it.

D Additional statistics of the test sets

Table 11 shows the type-token ratios for the source and target side of each of the test sets, shown for the four main domains. As mentioned previously, texts are tokenised using the language-specific Spacy models (Honnibal and Montani, 2017) where available. For Hebrew, we use the multilingual Spacy model as no language-specific model is available. The type-token ratio is calculated as the number of unique tokens divided by the total number of tokens. The absolute value depends not only on the lexical diversity of the text but also on the morphological complexity of the language in question.

	manuals		mastodon		news		user_review	
	src	trg	src	trg	src	trg	src	trg
From English								
en–cs	–	–	0.30	0.42	0.27	0.39	0.22	0.35
en–de	–	–	0.30	0.32	0.27	0.29	–	–
en–he	–	–	0.30	0.30	0.27	0.29	0.22	0.24
en–ja	–	–	0.30	0.23	0.27	0.19	0.22	0.17
en–ru	–	–	0.30	0.41	0.27	0.38	0.22	0.33
en–uk	–	–	0.30	0.41	0.27	0.38	0.22	0.34
en–zh	–	–	0.30	0.29	0.27	0.26	0.22	0.21
Other language directions								
cs–uk	–	–	–	–	0.43	0.41	–	–
de–en	0.32	0.23	0.49	0.42	0.34	0.26	–	–
he–en	–	–	–	–	0.34	0.09	–	–
ja–en	–	–	–	–	0.22	0.23	0.22	0.21
ru–en	0.47	0.28	–	–	0.40	0.24	–	–
uk–en	–	–	–	–	0.36	0.21	–	–
zh–en	0.25	0.25	–	–	0.23	0.19	0.22	0.17

Table 11: Type-token ratio for individual source languages used in the general translation test sets.

E News Task System Submission Summaries

This section lists all the submissions to the translation task and provides the authors' descriptions of their submission.

E.1 AIRC (Rikters and Miwa, 2023)

AIRC trained constrained track models for translation between English, German, and Japanese. Before training the final models we first filtered the parallel and monolingual data (Rikters, 2018), then performed iterative back-translation as well as parallel data distillation to be used for non-autoregressive model training. We experimented with training Transformer models, Mega (Ma et al., 2022) models, and custom non-autoregressive sequence-to-sequence models with encoder and decoder weights initialised by multilingual BERT base. Our primary submissions contain translations from ensembles of two Mega model checkpoints and our contrastive submissions are generated by our non-autoregressive models.

E.2 ANVITA (no associated paper)

ANVITA-ZhJa Machine Translation system for WMT2023 Shared Task:General MT(News). This paper describes ANVITA-ZhJa MT system, architected for submission to WMT 2023 General Machine Translation(News) shared task by the ANVITA team, where the team participated in 4 translation directions: Chinese, Japanese→English and English→Chinese, Japanese. ANVITA-ZhJa MT system comprised of four NMT models.Chinese, Japanese→English and English→Chinese, Japanese multilingual models for primary and Chinese→English and English→Chinese bilingual models for contrastive submissions. Base MT models are built using transformer(base) architecture, trained over the organizer provided parallel corpus and subsequently used deep transformer with added layers and other parameters. We also distilled corpus using heuristics based filtering and used model ensemble for enhanced performance.

E.3 CUNI-DocTransformer (Popel, 2020)

Exactly the same system as submitted in WMT20, document-level Transformer trained with Block Backtranslation.

E.4 CUNI-GA (Jon et al., 2023)

Our submission is a result of applying a novel n-best list reranking and modification method on translation candidates produced by two other competing systems, CUNI-Transformer and CUNI-DocTransformer. Our method uses a genetic algorithm and MBR decoding to search for optimal translation under a given metric (in our case, a weighted combination of ChrF, BLEU, COMET22-DA, and COMET22-QE-DA).

E.5 CUNI-Transformer (Popel, 2020)

The English↔Czech sentence-level models are exactly the same as submitted in WMT20 (Popel, 2020). The Ukrainian↔Czech models are very similar, also trained with Block Backtranslation.

E.6 GTCOM (Zong, 2023)

GTCOM uses transformer as the basic architecture and leverages multilingual models to improve translation quality. Besides, GTCOM does a lot of data cleaning and data augmentation work.

E.7 HW-TSC (Wu et al., 2023b)

HW-TSC's submission is a standard Transformer model equipped with our recent technique.

E.8 IOL-Research (Zhang, 2023)

This paper describes the IOL Research team's submission system for the WMT23 General Machine Translation shared task. We participate in two language translation directions, including English-to-Chinese and Chinese-to-English. Our final primary submissions belong to constrained systems, which means for both translation directions we only use officially provided monolingual and bilingual data to train the translation systems. Our systems are based on Transformer architecture with pre-norm or deep-norm, which has been proven to be helpful for training deeper models. We employ methods such as back-translation, data diversification, domain fine-tuning and model ensemble to build our translation systems. Another important aspect is that we carefully conduct data cleaning and use as much monolingual data as possible for data augmentation.

E.9 TeamKYB (LI et al., 2023)

We here describe our neural machine translation system for the general machine translation shared task in WMT 2023. Our systems are based on the Transformer with base settings. We trained our model with preprocessed train data. We collect multiple checkpoint from our model and performed inference with several hyperparameter settings. Collected translations were processed via some rule-based corrections. We chose best translation from the results by using N-best ranking method.

E.10 Lan-BridgeMT (Wu and Hu, 2023)

With the emergence of large-scale models, various industries have undergone significant transformations, particularly in the realm of document-level machine translation. This has introduced a novel research paradigm that we have embraced in our participation in the WMT23 competition. Focusing on advancements in models such as chatGPT and GPT4, we have undertaken numerous prompt-based experiments. Our objective is to achieve optimal human evaluation results for document-level machine translation, resulting in our submission of the final outcomes in the general track.

E.11 MUNI-NLP (Rychlý and Teslia, 2023)

MUNI-NLP system is a standard transformer.

E.12 NAIST-NICT (Deguchi et al., 2023)

In this paper, we describe our NAIST-NICT submission to the WMT'23 English-Japanese general machine translation task. Our system generates diverse translation candidates and reranks them with a two-stage reranking system to find the best translation. We first generate 50 candidates each from 18 different translation methods using a variety of techniques to increase the diversity of the translation candidates. We trained 7 different models per language direction using different combinations of hyperparameters. From these models we used various decoding algorithms, ensembling the models, and using kNN-MT. The 900 translation candidates go through a two-stage reranking system in order to find the most promising candidate.The first step compares the 50 candidates from each translation method using DrNMT and returns the one with the highest score. The final 18 candidates are ranked using COMET-MBR, and the highest scoring is returned as the system output. We found that generating diverse translation candidates improves the translation quality by using the well-designed relanker model.

E.13 PROMT (Molchanov and Kovalenko, 2023)

This paper describes the PROMT submissions for the WMT23 Shared General Translation Task. This year we participated in two directions of the Shared Translation Task: English to Russian and Russian to English. Our models are trained with the MarianNMT toolkit using the transformer-big configuration. We use BPE for text encoding, both models are unconstrained. We achieve competitive results according to automatic metrics in both directions.

E.14 SRPH (Cruz, 2023)

We submit single-model encode-decoder Transformer systems for the constrained English to Hebrew and Hebrew to English translation directions. Our dataset is cleaned and filtered via a combination of heuristic-based, ratio-based, and embedding-based (LaBSE) methods, resulting in a dataset with high alignment. We train models with heavy use of back-translation and decode using Noisy Channel Reranking using a reverse model and a language model trained with contest data.

E.15 SKIM (Kudo et al., 2023)

The SKIM team submission took a standard procedure of building ensemble Transformer models, including base-model training, data augmentation using back-translation of base models, and retraining several final models using back-translated training data. Each final model has its own architecture and configuration, including a 10.5B parameter at most, substituting self and cross sublayers in decoder with cross+self-attention sub-layer (Peitz et al., 2019). We select the best candidate from large candidate pools, namely 70 translations generated from 16 distinct models for each sentence, with an MBR reranking method using COMET and COMET-QE (Fernandes et al., 2022). We also applied data augmentation and selection techniques to training data of the Transformer models.

E.16 UPCite-CLILLF (no associated paper)

In this biomedical shared task, we have created data filters to better "choose" relevant training data for fine-tuning, among provided training data sources. In particular, we have used the textometric analysis tool ITRAMEUR to filter the segments and terms that characterize the test set and then extracted them from training data to fine-tune MBart-50 baseline (`decoder_attention_heads: 16`, `decoder_ffn_dim: 4096`, `decoder_layers: 12`, `encoder_attention_heads: 16`, `encoder_ffn_dim: 4096`, `encoder_layers: 12`, `num_hidden_layers: 12`, `max_length: 200`, `epoch: 3`). In doing so, we hope to meet several objectives : to build feasible fine-tuning strategy to train biomedical in-domain fr<->en models ; to specify filtering criteria of in-domain training data and to compare models' predictions, fine-tuning data and test set in order to better understand how neural machine translation systems work. We will also compare the pipeline of the shared task of this year to those of the past 2 years to evaluate the benefits of our training strategies of in-domain machine translation models.

E.17 UvA-LTL (Wu et al., 2023a)

We present our WMT system, UvA-MT, in the WMT 2023 shared general translation task. This year, we developed a single Multilingual Machine Translation (MMT) system to participate in the two-directional translation track between English and Hebrew. The main architecture is based on the prior work of Beyond Shared Vocabulary (Wu and Monz, 2023). We scaled it up to a transformer-large level (422M parameters). Additionally, we employed back translation to generate synthetic data and labeled them with a new language tag. After convergence, we further fine-tuned the system without using synthetic data. Several domain shift techniques were also introduced, such as the domain-aware language model, to filter monolingual data.

E.18 YiShu (Min et al., 2023)

Yishu's team participated in WMT23 Machine Translation Competition and adopted the most advanced neural machine translation method. They use Transformer model structure and use large-scale parallel corpus for training. In order to improve the translation quality, the team adopted cutting-edge data preprocessing technology, various attention mechanisms and improved decoding strategies. In addition, they also carried out in-depth parameter adjustment and model optimization. Yishu team incorporated evaluation indicators such as BLEU and TER into the training constraints of the model to achieve better translation performance. They strive for high accuracy and fluency in the competition, and strive to achieve excellent results in the field of translation.

E.19 LanguageX (Zeng, 2023)

LanguageX's submission is a many-to-many encoder decoder transformer model.

F Automatic scores

This section contains automatic metric scores. While human judgement is the official ranking of systems and their performance, we share automatic scores to show expected system performance for various testsets.

We use COMET (Rei et al., 2020) as the primary metric and chrF (Popović, 2015) as the secondary metric, following recommendation by (Kocmi et al., 2021). We also present BLEU (Papineni et al., 2002) scores as it is still a widely used metric. The COMET scores are calculated with the default model `Unbabel/wmt22-comet-da`. The chrF and BLEU scores are calculated using SacreBLEU (Post, 2018). Scores are multiplied by 100. We ranked the systems according to their scores. Unconstrained systems are indicated with a grey background in the tables.

System	COMET
CUNI-GA	90.9
GPT4-5shot	90.8
ONLINE-W	89.4
GTCOM_Peter	88.9
ONLINE-B	88.8
ONLINE-A	88.2
CUNI-Transformer	88.0
ONLINE-G	87.7
MUNI-NLP	87.0
ONLINE-Y	86.5
NLLB_Greedy	86.3
NLLB_MBR_BLEU	86.3
Lan-BridgeMT	86.0

System	chrF
GPT4-5shot	61.0
CUNI-GA	57.9
GTCOM_Peter	57.6
CUNI-Transformer	57.4
MUNI-NLP	57.0
Lan-BridgeMT	55.7
ONLINE-W	55.0
ONLINE-B	54.7
ONLINE-A	54.4
ONLINE-G	53.7
ONLINE-Y	53.4
NLLB_Greedy	52.5
NLLB_MBR_BLEU	52.3

System	BLEU
GPT4-5shot	32.8
CUNI-Transformer	30.2
GTCOM_Peter	29.8
CUNI-GA	29.5
MUNI-NLP	28.3
Lan-BridgeMT	27.5
ONLINE-W	26.8
ONLINE-B	25.7
ONLINE-A	25.4
NLLB_MBR_BLEU	25.1
NLLB_Greedy	24.9
ONLINE-G	24.8
ONLINE-Y	24.2

Table 12: Scores for the cs→uk translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	COMET
ONLINE-W	91.8
CUNI-GA	90.8
ONLINE-B	89.9
GPT4-5shot	89.4
ONLINE-A	88.4
CUNI-DocTransformer	88.3
GTCOM_Peter	87.7
ONLINE-M	87.4
Lan-BridgeMT	87.3
CUNI-Transformer	87.2
NLLB_Greedy	87.1
ONLINE-Y	87.0
NLLB_MBR_BLEU	86.9
ONLINE-G	85.9
ZengHuiMT	85.4

System	chrF
ONLINE-W	76.3
ONLINE-B	70.4
ZengHuiMT	67.5
ONLINE-A	66.3
CUNI-GA	65.9
GTCOM_Peter	65.4
CUNI-DocTransformer	65.1
ONLINE-Y	64.6
CUNI-Transformer	63.9
Lan-BridgeMT	63.8
ONLINE-G	63.7
ONLINE-M	63.2
GPT4-5shot	62.3
NLLB_Greedy	60.0
NLLB_MBR_BLEU	59.1

System	BLEU
ONLINE-W	59.4
ONLINE-B	50.1
ONLINE-A	43.4
CUNI-GA	43.3
ZengHuiMT	43.1
CUNI-DocTransformer	42.5
GTCOM_Peter	42.3
CUNI-Transformer	41.4
ONLINE-Y	40.8
Lan-BridgeMT	40.7
ONLINE-G	39.6
ONLINE-M	39.6
GPT4-5shot	37.8
NLLB_Greedy	35.9
NLLB_MBR_BLEU	35.1

Table 13: Scores for the en→cs translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	COMET
GPT4-5shot	86.3
ONLINE-W	86.0
ONLINE-B	85.6
ONLINE-A	85.5
ONLINE-Y	84.9
ONLINE-M	84.8
ONLINE-G	84.6
GTCOM_Peter	82.7
NLLB_MBR_BLEU	81.4
ZengHuiMT	81.1
Lan-BridgeMT	80.9
NLLB_Greedy	79.9
AIRC	78.7

System	chrF
ONLINE-W	72.1
ONLINE-A	70.0
GPT4-5shot	69.8
ONLINE-B	69.1
ONLINE-G	69.1
ONLINE-Y	68.4
ZengHuiMT	67.6
Lan-BridgeMT	66.7
GTCOM_Peter	66.6
ONLINE-M	66.5
NLLB_MBR_BLEU	57.6
NLLB_Greedy	57.3
AIRC	57.2

System	BLEU
ONLINE-W	51.8
GPT4-5shot	47.9
ONLINE-A	47.9
ONLINE-B	46.3
ONLINE-G	46.0
ONLINE-Y	43.9
GTCOM_Peter	42.2
Lan-BridgeMT	42.1
ONLINE-M	41.3
ZengHuiMT	40.8
NLLB_Greedy	33.1
AIRC	32.4
NLLB_MBR_BLEU	32.4

Table 14: Scores for the de→en translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	COMET
ONLINE-W	85.5
GPT4-5shot	85.0
ONLINE-B	84.8
ONLINE-Y	84.1
ONLINE-A	83.7
ONLINE-G	82.5
ONLINE-M	81.7
Lan-BridgeMT	80.4
ZengHuiMT	79.4
NLLB_MBR_BLEU	78.0
NLLB_Greedy	77.9
AIRC	72.9

System	chrF
ONLINE-W	71.8
ONLINE-A	69.7
ZengHuiMT	69.4
GPT4-5shot	69.1
ONLINE-B	69.1
ONLINE-Y	69.1
ONLINE-G	69.0
ONLINE-M	66.9
Lan-BridgeMT	66.1
NLLB_Greedy	56.2
NLLB_MBR_BLEU	55.4
AIRC	52.2

System	BLEU
ONLINE-W	47.8
ONLINE-A	43.7
GPT4-5shot	43.6
ONLINE-Y	43.6
ONLINE-G	43.2
ONLINE-B	42.7
ONLINE-M	40.5
ZengHuiMT	40.5
Lan-BridgeMT	39.4
NLLB_Greedy	31.1
NLLB_MBR_BLEU	29.6
AIRC	26.5

Table 15: Scores for the en→de translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	COMET
ONLINE-B	89.9
ONLINE-A	87.0
GPT4-5shot	86.9
GTCOM_Peter	86.7
ONLINE-G	85.6
ZengHuiMT	85.6
ONLINE-Y	84.9
UvA-LTL	84.7
NLLB_MBR_BLEU	82.9
NLLB_Greedy	82.8
Samsung_Research_Philippines	82.6
Lan-BridgeMT	82.4

System	chrF
ONLINE-B	87.5
ZengHuiMT	76.3
GTCOM_Peter	76.2
ONLINE-A	73.3
GPT4-5shot	71.4
UvA-LTL	70.9
ONLINE-Y	70.5
ONLINE-G	69.8
NLLB_Greedy	64.4
Lan-BridgeMT	63.5
NLLB_MBR_BLEU	63.0
Samsung_Research_Philippines	55.5

System	BLEU
ONLINE-B	76.5
GTCOM_Peter	59.2
ZengHuiMT	56.6
ONLINE-A	53.9
GPT4-5shot	51.2
UvA-LTL	51.0
ONLINE-Y	49.8
ONLINE-G	49.3
NLLB_Greedy	42.5
Lan-BridgeMT	41.4
NLLB_MBR_BLEU	40.7
Samsung_Research_Philippines	34.0

Table 16: Scores for the he→en (refA) translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6 |nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	COMET
GPT4-5shot	86.4
ONLINE-B	85.6
ONLINE-A	85.3
GTCOM_Peter	84.5
ONLINE-G	84.0
UvA-LTL	83.3
ZengHuiMT	83.3
ONLINE-Y	82.9
NLLB_MBR_BLEU	81.8
NLLB_Greedy	81.7
Lan-BridgeMT	81.3
Samsung_Research_Philippines	81.3

System	chrF
GPT4-5shot	69.5
ONLINE-B	66.5
ONLINE-A	65.6
GTCOM_Peter	65.3
ZengHuiMT	65.1
UvA-LTL	63.3
ONLINE-G	62.8
ONLINE-Y	62.0
NLLB_Greedy	59.6
Lan-BridgeMT	59.0
NLLB_MBR_BLEU	58.6
Samsung_Research_Philippines	51.3

System	BLEU
GPT4-5shot	50.4
ONLINE-B	45.0
GTCOM_Peter	44.4
ONLINE-A	44.4
UvA-LTL	41.7
ZengHuiMT	41.7
ONLINE-G	40.9
ONLINE-Y	38.5
NLLB_Greedy	37.1
Lan-BridgeMT	36.2
NLLB_MBR_BLEU	36.2
Samsung_Research_Philippines	29.8

Table 17: Scores for the he→en (refB) translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6 |nw:0|space:no|version:2.3.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1), COMET (Unbabel/wmt22-comet-da).

System	COMET
ONLINE-B	86.4
ONLINE-A	85.7
GPT4-5shot	84.9
GTCOM_Peter	84.7
ONLINE-Y	84.7
UvA-LTL	84.2
Samsung_Research_Philippines	83.7
Lan-BridgeMT	83.0
NLLB_Greedy	82.9
ZengHuiMT	82.7
NLLB_MBR_BLEU	82.5
ONLINE-G	82.2

System	chrF
ONLINE-B	66.4
ZengHuiMT	62.1
ONLINE-A	61.7
GTCOM_Peter	61.1
ONLINE-Y	60.4
UvA-LTL	59.0
ONLINE-G	58.1
Samsung_Research_Philippines	57.3
Lan-BridgeMT	54.9
NLLB_Greedy	54.8
NLLB_MBR_BLEU	54.3
GPT4-5shot	54.0

System	BLEU
ONLINE-B	47.8
ONLINE-A	38.9
GTCOM_Peter	37.2
ONLINE-Y	37.2
ZengHuiMT	36.5
UvA-LTL	35.0
Samsung_Research_Philippines	33.3
ONLINE-G	33.2
NLLB_MBR_BLEU	30.8
Lan-BridgeMT	30.5
NLLB_Greedy	30.3
GPT4-5shot	27.0

Table 18: Scores for the en→he translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	COMET
SKIM	84.0
GPT4-5shot	83.4
ONLINE-W	82.3
NAIST-NICT	81.9
ONLINE-Y	81.6
ONLINE-B	81.5
ONLINE-A	81.0
GTCOM_Peter	80.2
ANVITA	79.5
Lan-BridgeMT	79.3
ZengHuiMT	79.2
ONLINE-G	77.8
ONLINE-M	77.5
KYB	76.6
NLLB_MBR_BLEU	75.2
AIRC	74.5
NLLB_Greedy	74.3

System	chrF
ONLINE-W	51.4
GPT4-5shot	51.2
SKIM	51.1
ONLINE-A	49.6
NAIST-NICT	49.5
ONLINE-Y	49.5
ZengHuiMT	49.5
ONLINE-B	49.3
GTCOM_Peter	48.7
Lan-BridgeMT	47.3
ANVITA	46.7
ONLINE-G	45.5
KYB	43.9
ONLINE-M	43.9
AIRC	40.5
NLLB_MBR_BLEU	39.2
NLLB_Greedy	39.0

System	BLEU
ONLINE-W	25.9
SKIM	24.8
GPT4-5shot	24.1
ONLINE-B	23.9
NAIST-NICT	23.0
ONLINE-A	23.0
ZengHuiMT	22.6
GTCOM_Peter	22.3
ONLINE-Y	22.3
ANVITA	20.9
Lan-BridgeMT	20.2
ONLINE-G	18.3
KYB	17.6
ONLINE-M	17.2
AIRC	14.9
NLLB_MBR_BLEU	14.7
NLLB_Greedy	14.2

Table 19: Scores for the ja→en translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	COMET
ONLINE-B	88.2
ONLINE-W	87.5
ONLINE-Y	87.3
GPT4-5shot	87.0
SKIM	86.6
NAIST-NICT	86.2
ZengHuiMT	85.3
ONLINE-A	85.2
Lan-BridgeMT	84.5
ONLINE-M	13.3
ANVITA	82.7
KYB	80.8
AIRC	80.7
ONLINE-G	80.4
NLLB_Greedy	79.3
NLLB_MBR_BLEU	77.7

System	chrF
ONLINE-B	35.2
ONLINE-Y	34.1
ONLINE-W	33.5
SKIM	33.5
ZengHuiMT	32.9
NAIST-NICT	32.0
ONLINE-A	31.4
GPT4-5shot	31.0
Lan-BridgeMT	30.4
ONLINE-M	29.6
ANVITA	29.3
KYB	27.7
AIRC	27.6
ONLINE-G	27.3
NLLB_Greedy	20.9
NLLB_MBR_BLEU	18.7

System	BLEU
ONLINE-B	25.3
ONLINE-W	24.5
ONLINE-Y	24.5
SKIM	24.3
NAIST-NICT	22.6
ZengHuiMT	22.6
ONLINE-A	21.4
GPT4-5shot	21.3
Lan-BridgeMT	20.5
ONLINE-M	19.8
ANVITA	19.4
KYB	17.8
AIRC	17.6
ONLINE-G	17.2
NLLB_Greedy	11.3
NLLB_MBR_BLEU	9.0

Table 20: Scores for the en→ja translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:ja-mecab-0.996-IPA|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	COMET
GPT4-5shot	83.5
ONLINE-Y	82.5
ONLINE-B	82.3
ONLINE-W	82.2
ONLINE-G	82.0
ONLINE-A	81.9
PROMT	80.9
ONLINE-M	80.7
NLLB_MBR_BLEU	80.5
NLLB_Greedy	80.1
Lan-BridgeMT	79.9
ZengHuiMT	79.5

System	chrF
GPT4-5shot	60.4
ONLINE-G	59.6
ONLINE-A	59.4
ONLINE-B	59.4
ZengHuiMT	58.9
ONLINE-Y	58.6
PROMT	58.4
ONLINE-W	58.3
Lan-BridgeMT	57.4
ONLINE-M	56.7
NLLB_MBR_BLEU	55.8
NLLB_Greedy	55.5

System	BLEU
ONLINE-B	34.5
GPT4-5shot	34.4
ONLINE-G	34.0
ONLINE-A	33.8
ONLINE-Y	33.2
ONLINE-W	33.1
PROMT	32.8
Lan-BridgeMT	31.8
ZengHuiMT	31.3
NLLB_MBR_BLEU	31.0
ONLINE-M	30.7
NLLB_Greedy	30.3

Table 21: Scores for the ru→en translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	COMET
ONLINE-G	86.6
ONLINE-W	86.6
ONLINE-B	86.2
GPT4-5shot	86.1
ONLINE-Y	85.5
ONLINE-A	85.3
ONLINE-M	83.2
Lan-BridgeMT	83.1
NLLB_Greedy	82.9
NLLB_MBR_BLEU	82.7
PROMT	82.3
ZengHuiMT	81.3

System	chrF
ONLINE-B	61.9
ONLINE-A	59.0
ONLINE-G	58.9
ZengHuiMT	58.8
ONLINE-W	56.6
ONLINE-Y	56.4
GPT4-5shot	56.2
Lan-BridgeMT	55.7
PROMT	55.4
ONLINE-M	55.1
NLLB_Greedy	53.3
NLLB_MBR_BLEU	53.1

System	BLEU
ONLINE-B	40.4
ONLINE-A	34.8
ONLINE-G	32.9
ONLINE-Y	32.0
ZengHuiMT	31.6
ONLINE-W	31.4
ONLINE-M	30.9
Lan-BridgeMT	30.7
GPT4-5shot	30.6
PROMT	30.5
NLLB_MBR_BLEU	28.4
NLLB_Greedy	28.2

Table 22: Scores for the en→ru translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	COMET
ONLINE-W	87.5
GPT4-5shot	87.1
ONLINE-B	86.8
GTCOM_Peter	86.3
ONLINE-A	86.3
ONLINE-G	86.2
ONLINE-Y	85.8
Lan-BridgeMT	84.8
ZengHuiMT	84.4
NLLB_MBR_BLEU	84.3
NLLB_Greedy	84.2

System	chrF
GTCOM_Peter	69.3
ONLINE-W	69.2
ONLINE-B	69.0
ZengHuiMT	68.5
ONLINE-A	68.3
ONLINE-Y	68.2
GPT4-5shot	68.1
ONLINE-G	68.0
Lan-BridgeMT	66.2
NLLB_Greedy	62.4
NLLB_MBR_BLEU	62.4

System	BLEU
ONLINE-W	47.4
GTCOM_Peter	46.4
ONLINE-B	46.0
ONLINE-A	45.9
ONLINE-Y	45.7
ONLINE-G	44.9
GPT4-5shot	43.9
ZengHuiMT	43.5
Lan-BridgeMT	42.3
NLLB_MBR_BLEU	38.1
NLLB_Greedy	37.8

Table 23: Scores for the uk→en translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	COMET
ONLINE-W	86.7
ONLINE-B	85.6
GPT4-5shot	85.3
ONLINE-G	85.3
ONLINE-A	83.2
ONLINE-Y	82.9
GTCOM_Peter	82.1
NLLB_Greedy	82.1
NLLB_MBR_BLEU	81.7
Lan-BridgeMT	80.4
ZengHuiMT	79.0

System	chrF
ONLINE-B	61.7
ONLINE-W	59.2
ZengHuiMT	56.4
ONLINE-G	56.1
ONLINE-A	55.8
ONLINE-Y	55.4
GTCOM_Peter	54.4
GPT4-5shot	53.0
Lan-BridgeMT	51.9
NLLB_Greedy	50.8
NLLB_MBR_BLEU	50.5

System	BLEU
ONLINE-B	39.8
ONLINE-W	34.9
ONLINE-A	30.3
ONLINE-Y	29.5
ONLINE-G	28.6
ZengHuiMT	27.8
GTCOM_Peter	27.5
GPT4-5shot	25.2
NLLB_MBR_BLEU	24.9
Lan-BridgeMT	24.6
NLLB_Greedy	24.5

Table 24: Scores for the en→uk translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	COMET
HW-TSC	82.8
ONLINE-B	82.7
Yishu	82.7
GPT4-5shot	81.6
Lan-BridgeMT	81.2
ONLINE-G	80.9
ONLINE-Y	80.6
ONLINE-A	80.3
ZengHuiMT	79.6
ONLINE-W	79.3
IOL_Research	79.2
ONLINE-M	77.7
NLLB_MBR_BLEU	76.8
ANVITA	76.6
NLLB_Greedy	76.4

System	chrF
HW-TSC	57.5
ONLINE-B	57.5
Yishu	57.4
ZengHuiMT	54.6
ONLINE-G	53.9
ONLINE-A	53.4
GPT4-5shot	53.1
Lan-BridgeMT	53.1
ONLINE-W	52.5
IOL_Research	52.4
ONLINE-Y	52.3
ONLINE-M	49.7
ANVITA	47.1
NLLB_Greedy	46.1
NLLB_MBR_BLEU	45.8

System	BLEU
HW-TSC	33.6
ONLINE-B	33.5
Yishu	33.4
ONLINE-A	28.3
Lan-BridgeMT	27.3
IOL_Research	27.2
ZengHuiMT	27.0
GPT4-5shot	26.8
ONLINE-G	26.6
ONLINE-W	26.4
ONLINE-Y	25.0
ONLINE-M	23.5
ANVITA	21.8
NLLB_Greedy	20.5
NLLB_MBR_BLEU	19.8

Table 25: Scores for the zh→en translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

System	COMET
ONLINE-B	88.1
Yishu	88.1
HW-TSC	87.3
GPT4-5shot	87.1
ONLINE-W	86.8
Lan-BridgeMT	86.6
ONLINE-Y	86.5
ONLINE-A	86.2
IOL_Research	85.3
ZengHuiMT	84.3
ONLINE-M	84.2
ONLINE-G	83.8
NLLB_Greedy	75.7
ANVITA	75.6
NLLB_MBR_BLEU	71.5

System	chrF
HW-TSC	53.8
Yishu	53.0
ONLINE-B	52.9
ONLINE-A	52.8
IOL_Research	51.9
ONLINE-M	50.6
ONLINE-Y	49.8
ONLINE-G	49.4
ONLINE-W	47.3
ZengHuiMT	47.0
Lan-BridgeMT	46.8
GPT4-5shot	46.5
ANVITA	36.9
NLLB_Greedy	26.3
NLLB_MBR_BLEU	21.1

System	BLEU
HW-TSC	58.6
ONLINE-A	58.5
Yishu	57.6
ONLINE-B	57.5
IOL_Research	56.9
ONLINE-M	54.9
ONLINE-Y	54.2
ONLINE-G	54.1
ZengHuiMT	52.9
ONLINE-W	52.1
Lan-BridgeMT	50.2
GPT4-5shot	49.6
ANVITA	38.9
NLLB_Greedy	27.4
NLLB_MBR_BLEU	19.1

Table 26: Scores for the en→zh translation task: chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), BLEU (nrefs:1|case:mixed|eff:no|tok:zh|smooth:exp|version:2.2.1), COMET (Unbabel/wmt22-comet-da).

G Head to head comparisons

Following tables show differences in average human scores for each language pair. The numbers in each of the tables' cells indicate the difference in average human scores for the system in that column and the system in that row.

Because there were so many systems and data conditions the significance of each pairwise comparison needs to be quantified. We applied Wilcoxon rank-sum test to measure the likelihood that such differences could occur simply by chance. In the following tables $\star$ indicates statistical significance at $p < 0.05$, † indicates statistical significance at $p < 0.01$, and ‡ indicates statistical significance at $p < 0.001$, according to Wilcoxon rank-sum test.

Each table contains final rows showing the average score achieved by that system and the rank range according to Wilcoxon rank-sum test ($p < 0.05$). Gray lines separate clusters based on non-overlapping rank ranges.

Czech→Ukrainian

	ONLINE-B	GPT4-5shot	Human-refA	ONLINE-W	CUNI-GA	CUNI-Transformer	GTCOM_DLUT	ONLINE-A	ONLINE-G	ONLINE-Y	MUNI-NLP	Lan-BridgeMT	NLLB_MBR_BLEU	NLLB_Greedy
ONLINE-B	—	0.1	0.4	0.9⋆	1.3‡	1.8†	2.4⋆	3.1⋆	4.1‡	5.0‡	5.0‡	6.2‡	6.7‡	7.0‡
GPT4-5shot	-0.1	—	0.4	0.8†	1.2‡	1.8‡	2.3†	3.1‡	4.1‡	4.9‡	4.9‡	6.2‡	6.7‡	6.9‡
Human-refA	-0.4	-0.4	—	0.5‡	0.9‡	1.4‡	1.9‡	2.7‡	3.7‡	4.6‡	4.6‡	5.8‡	6.3‡	6.6‡
ONLINE-W	-0.9	-0.8	-0.5	—	0.4	0.9	1.5	2.2	3.2‡	4.1‡	4.1‡	5.3‡	5.8‡	6.1‡
CUNI-GA	-1.3	-1.2	-0.9	-0.4	—	0.6	1.1	1.8	2.9†	3.7‡	3.7†	5.0‡	5.5‡	5.7‡
CUNI-Transformer	-1.8	-1.8	-1.4	-0.9	-0.6	—	0.5	1.3	2.3†	3.1‡	3.2‡	4.4‡	4.9‡	5.1‡
GTCOM_DLUT	-2.4	-2.3	-1.9	-1.5	-1.1	-0.5	—	0.8	1.8‡	2.6‡	2.6‡	3.9‡	4.4‡	4.6‡
ONLINE-A	-3.1	-3.1	-2.7	-2.2	-1.8	-1.3	-0.8	—	1.0‡	1.9‡	1.9‡	3.1‡	3.6‡	3.9‡
ONLINE-G	-4.1	-4.1	-3.7	-3.2	-2.9	-2.3	-1.8	-1.0	—	0.8	0.9	2.1⋆	2.6†	2.8‡
ONLINE-Y	-5.0	-4.9	-4.6	-4.1	-3.7	-3.1	-2.6	-1.9	-0.8	—	0.0	1.3	1.8	2.0†
MUNI-NLP	-5.0	-4.9	-4.6	-4.1	-3.7	-3.2	-2.6	-1.9	-0.9	-0.0	—	1.2	1.7	2.0‡
Lan-BridgeMT	-6.2	-6.2	-5.8	-5.3	-5.0	-4.4	-3.9	-3.1	-2.1	-1.3	-1.2	—	0.5	0.7⋆
NLLB_MBR_BLEU	-6.7	-6.7	-6.3	-5.8	-5.5	-4.9	-4.4	-3.6	-2.6	-1.8	-1.7	-0.5	—	0.2⋆
NLLB_Greedy	-7.0	-6.9	-6.6	-6.1	-5.7	-5.1	-4.6	-3.9	-2.8	-2.0	-2.0	-0.7	-0.2	—
score	83.7	83.6	83.2	82.8	82.4	81.8	81.3	80.6	79.5	78.7	78.7	77.4	76.9	76.7
rank	1-3	1-3	1-3	4-8	4-8	4-8	4-8	4-8	9-11	9-13	9-13	10-13	10-13	14

Table 27: Head to head comparison for Czech→Ukrainian systems

German→English

	GPT4-5shot	Human-refA	ONLINE-A	ONLINE-B	ONLINE-W	ONLINE-Y	ONLINE-G	GTCOM_DLUT	ONLINE-M	LanguageX	Lan-BridgeMT	NLLB_MBR_BLEU	AIRC	NLLB_Greedy
GPT4-5shot	—	0.4	0.8	1.2†	1.5†	2.3†	2.6‡	3.8‡	5.0‡	8.5‡	10.3‡	10.7‡	11.5‡	12.4‡
Human-refA	-0.4	—	0.4	0.8⋆	1.1⋆	1.9†	2.2‡	3.4‡	4.6‡	8.1‡	9.9†	10.3‡	11.1‡	12.0‡
ONLINE-A	-0.8	-0.4	—	0.4	0.7	1.6⋆	1.9†	3.0‡	4.2‡	7.7‡	9.6‡	9.9‡	10.8‡	11.7‡
ONLINE-B	-1.2	-0.8	-0.4	—	0.3	1.1	1.4⋆	2.6‡	3.8‡	7.3‡	9.2‡	9.5‡	10.3‡	11.2‡
ONLINE-W	-1.5	-1.1	-0.7	-0.3	—	0.8	1.1⋆	2.3‡	3.5‡	7.0‡	8.9‡	9.2‡	10.0‡	10.9‡
ONLINE-Y	-2.3	-1.9	-1.6	-1.1	-0.8	—	0.3	1.5‡	2.7†	6.2‡	8.0‡	8.4‡	9.2‡	10.1‡
ONLINE-G	-2.6	-2.2	-1.9	-1.4	-1.1	-0.3	—	1.2‡	2.4	5.9‡	7.7‡	8.1‡	8.9‡	9.8‡
GTCOM_DLUT	-3.8	-3.4	-3.0	-2.6	-2.3	-1.5	-1.2	—	1.2	4.7‡	6.6‡	6.9‡	7.8‡	8.6‡
ONLINE-M	-5.0	-4.6	-4.2	-3.8	-3.5	-2.7	-2.4	-1.2	—	3.5‡	5.3‡	5.7‡	6.5‡	7.4‡
LanguageX	-8.5	-8.1	-7.7	-7.3	-7.0	-6.2	-5.9	-4.7	-3.5	—	1.9	2.2†	3.0‡	3.9†
Lan-BridgeMT	-10.3	-9.9	-9.6	-9.2	-8.9	-8.0	-7.7	-6.6	-5.3	-1.9	—	0.3	1.2⋆	2.1
NLLB_MBR_BLEU	-10.7	-10.3	-9.9	-9.5	-9.2	-8.4	-8.1	-6.9	-5.7	-2.2	-0.3	—	0.8	1.7
AIRC	-11.5	-11.1	-10.8	-10.3	-10.0	-9.2	-8.9	-7.8	-6.5	-3.0	-1.2	-0.8	—	0.9
NLLB_Greedy	-12.4	-12.0	-11.7	-11.2	-10.9	-10.1	-9.8	-8.6	-7.4	-3.9	-2.1	-1.7	-0.9	—
score	90.3	89.9	89.6	89.1	88.8	88.0	87.7	86.5	85.3	81.8	80.0	79.6	78.8	77.9
rank	1-3	1-3	1-5	3-6	3-6	4-7	6-8	8-9	7-9	10-11	10-13	11-14	12-14	11-14

Table 28: Head to head comparison for German→English systems

English→Czech

	Human-refA	ONLINE-W	GPT4-5shot	CUNI-GA	ONLINE-A	CUNI-DocTransformer	ONLINE-B	NLLB_MBR_BLEU	GTCOM_DLUT	CUNI-Transformer	NLLB_Greedy	ONLINE-M	ONLINE-G	ONLINE-Y	Lan-BridgeMT	LanguageX
Human-refA	—	1.3⋆	3.6‡	5.0‡	5.1‡	6.0‡	6.6‡	6.8‡	7.0‡	8.0‡	8.6‡	9.7‡	10.2‡	10.4‡	10.4‡	11.3‡
ONLINE-W	-1.3	—	2.3‡	3.7‡	3.8‡	4.7‡	5.3‡	5.5‡	5.7‡	6.7‡	7.3‡	8.4‡	8.9‡	9.1‡	9.1‡	10.0‡
GPT4-5shot	-3.6	-2.3	—	1.4	1.5†	2.5⋆	3.0	3.2‡	3.4†	4.4‡	5.1‡	6.1‡	6.6‡	6.8‡	6.8‡	7.7‡
CUNI-GA	-5.0	-3.7	-1.4	—	0.0‡	1.0‡	1.5⋆	1.8‡	2.0‡	3.0‡	3.6‡	4.7‡	5.1‡	5.3‡	5.4‡	6.3‡
ONLINE-A	-5.1	-3.8	-1.5	-0.0	—	1.0	1.5	1.7‡	1.9	2.9⋆	3.6‡	4.7†	5.1‡	5.3‡	5.4⋆	6.3‡
CUNI-DocTransformer	-6.0	-4.7	-2.5	-1.0	-1.0	—	0.5	0.7‡	0.9	1.9†	2.6‡	3.7‡	4.1‡	4.3‡	4.4†	5.3‡
ONLINE-B	-6.6	-5.3	-3.0	-1.5	-1.5	-0.5	—	0.2‡	0.4⋆	1.4‡	2.1‡	3.2‡	3.6‡	3.8‡	3.9‡	4.8‡
NLLB_MBR_BLEU	-6.8	-5.5	-3.2	-1.8	-1.7	-0.7	-0.2	—	0.2	1.2	1.9	3.0	3.4	3.6⋆	3.7	4.5‡
GTCOM_DLUT	-7.0	-5.7	-3.4	-2.0	-1.9	-0.9	-0.4	-0.2	—	1.0	1.7‡	2.8†	3.2‡	3.4‡	3.5	4.3‡
CUNI-Transformer	-8.0	-6.7	-4.4	-3.0	-2.9	-1.9	-1.4	-1.2	-1.0	—	0.7⋆	1.7	2.2†	2.4‡	2.4	3.3‡
NLLB_Greedy	-8.6	-7.3	-5.1	-3.6	-3.6	-2.6	-2.1	-1.9	-1.7	-0.7	—	1.1	1.5	1.7⋆	1.8	2.7‡
ONLINE-M	-9.7	-8.4	-6.1	-4.7	-4.7	-3.7	-3.2	-3.0	-2.8	-1.7	-1.1	—	0.4	0.6†	0.7	1.6‡
ONLINE-G	-10.2	-8.9	-6.6	-5.1	-5.1	-4.1	-3.6	-3.4	-3.2	-2.2	-1.5	-0.4	—	0.2	0.3	1.1†
ONLINE-Y	-10.4	-9.1	-6.8	-5.3	-5.3	-4.3	-3.8	-3.6	-3.4	-2.4	-1.7	-0.6	-0.2	—	0.1	1.0⋆
Lan-BridgeMT	-10.4	-9.1	-6.8	-5.4	-5.4	-4.4	-3.9	-3.7	-3.5	-2.4	-1.8	-0.7	-0.3	-0.1	—	0.9‡
LanguageX	-11.3	-10.0	-7.7	-6.3	-6.3	-5.3	-4.8	-4.5	-4.3	-3.3	-2.7	-1.6	-1.1	-1.0	-0.9	—
score	85.4	84.1	81.8	80.4	80.3	79.4	78.8	78.6	78.4	77.4	76.8	75.7	75.2	75.0	75.0	74.1
rank	1	2	3-5	3-4	5-8	5-8	4-7	8-14	6-11	8-12	10-14	9-14	10-15	13-15	8-15	16

Table 29: Head to head comparison for English→Czech systems

English→German

	GPT4-5shot	ONLINE-B	ONLINE-W	ONLINE-A	ONLINE-Y	Human-refA	ONLINE-M	ONLINE-G	Lan-BridgeMT	LanguageX	NLLB_MBR_BLEU	NLLB_Greedy	AIRC
GPT4-5shot	—	0.1	0.7	0.8	1.0†	1.3	2.3‡	3.4‡	5.0‡	6.3‡	12.1‡	13.2‡	15.4‡
ONLINE-B	-0.1	—	0.6	0.7	0.8⋆	1.2	2.2‡	3.3‡	4.8‡	6.2‡	12.0‡	13.1‡	15.2‡
ONLINE-W	-0.7	-0.6	—	0.2⋆	0.3‡	0.6	1.6‡	2.7‡	4.3‡	5.6‡	11.5‡	12.5‡	14.7‡
ONLINE-A	-0.8	-0.7	-0.2	—	0.1	0.5	1.4‡	2.6‡	4.1‡	5.5‡	11.3‡	12.4‡	14.5‡
ONLINE-Y	-1.0	-0.8	-0.3	-0.1	—	0.3	1.3†	2.5⋆	4.0‡	5.3‡	11.2‡	12.3‡	14.4‡
Human-refA	-1.3	-1.2	-0.6	-0.5	-0.3	—	1.0‡	2.1‡	3.7‡	5.0‡	10.8‡	11.9‡	14.1‡
ONLINE-M	-2.3	-2.2	-1.6	-1.4	-1.3	-1.0	—	1.1	2.7†	4.0‡	9.9‡	10.9‡	13.1‡
ONLINE-G	-3.4	-3.3	-2.7	-2.6	-2.5	-2.1	-1.1	—	1.5‡	2.9‡	8.7‡	9.8‡	11.9‡
Lan-BridgeMT	-5.0	-4.8	-4.3	-4.1	-4.0	-3.7	-2.7	-1.5	—	1.4⋆	7.2‡	8.3‡	10.4‡
LanguageX	-6.3	-6.2	-5.6	-5.5	-5.3	-5.0	-4.0	-2.9	-1.4	—	5.8‡	6.9‡	9.1‡
NLLB_MBR_BLEU	-12.1	-12.0	-11.5	-11.3	-11.2	-10.8	-9.9	-8.7	-7.2	-5.8	—	1.1	3.2‡
NLLB_Greedy	-13.2	-13.1	-12.5	-12.4	-12.3	-11.9	-10.9	-9.8	-8.3	-6.9	-1.1	—	2.2‡
AIRC	-15.4	-15.2	-14.7	-14.5	-14.4	-14.1	-13.1	-11.9	-10.4	-9.1	-3.2	-2.2	—
score	89.0	88.8	88.3	88.1	88.0	87.7	86.7	85.5	84.0	82.7	76.8	75.7	73.6
rank	1-5	1-5	1-4	2-6	4-6	1-6	7-8	7-8	9	10	11-12	11-12	13

Table 30: Head to head comparison for English→German systems

English→Japanese

	Human-refA	GPT4-5shot	ONLINE-B	ONLINE-Y	SKIM	ONLINE-W	LanguageX	ONLINE-A	NAIST-NICT	Lan-BridgeMT	ANVITA	ONLINE-M	KYB	AIRC	ONLINE-G	NLLB_Greedy	NLLB_MBR_BLEU
Human-refA	—	1.2‡	1.9	2.1†	2.2⋆	2.3‡	4.1‡	4.5‡	4.6‡	5.5‡	7.6‡	8.1‡	9.9‡	11.1‡	11.1‡	16.2‡	19.4‡
GPT4-5shot	-1.2	—	0.7	0.9	1.0	1.1	2.9⋆	3.3†	3.4†	4.3‡	6.4‡	6.9‡	8.8‡	9.9‡	10.0‡	15.0‡	18.3‡
ONLINE-B	-1.9	-0.7	—	0.2	0.3	0.4⋆	2.3‡	2.7‡	2.7‡	3.6‡	5.7‡	6.2‡	8.1‡	9.3‡	9.3‡	14.3‡	17.6‡
ONLINE-Y	-2.1	-0.9	-0.2	—	0.1	0.2	2.0‡	2.4‡	2.5‡	3.4‡	5.5‡	6.0‡	7.8‡	9.0‡	9.0‡	14.1‡	17.3‡
SKIM	-2.2	-1.0	-0.3	-0.1	—	0.1⋆	1.9‡	2.3‡	2.4‡	3.3‡	5.4‡	5.9‡	7.7‡	8.9‡	8.9‡	13.9‡	17.2‡
ONLINE-W	-2.3	-1.1	-0.4	-0.2	-0.1	—	1.8†	2.2†	2.3‡	3.2‡	5.3‡	5.8‡	7.6‡	8.8‡	8.8‡	13.8‡	17.1‡
LanguageX	-4.1	-2.9	-2.3	-2.0	-1.9	-1.8	—	0.4	0.5	1.4	3.5‡	4.0‡	5.8‡	7.0‡	7.0‡	12.0‡	15.3‡
ONLINE-A	-4.5	-3.3	-2.7	-2.4	-2.3	-2.2	-0.4	—	0.0	1.0	3.1‡	3.5‡	5.4‡	6.6‡	6.6‡	11.6‡	14.9‡
NAIST-NICT	-4.6	-3.4	-2.7	-2.5	-2.4	-2.3	-0.5	-0.0	—	0.9	3.0‡	3.5‡	5.4‡	6.5‡	6.6‡	11.6‡	14.9‡
Lan-BridgeMT	-5.5	-4.3	-3.6	-3.4	-3.3	-3.2	-1.4	-1.0	-0.9	—	2.1†	2.6‡	4.5‡	5.6‡	5.6‡	10.7‡	14.0‡
ANVITA	-7.6	-6.4	-5.7	-5.5	-5.4	-5.3	-3.5	-3.1	-3.0	-2.1	—	0.5	2.3‡	3.5‡	3.5‡	8.5‡	11.8‡
ONLINE-M	-8.1	-6.9	-6.2	-6.0	-5.9	-5.8	-4.0	-3.5	-3.5	-2.6	-0.5	—	1.9†	3.0‡	3.1†	8.1‡	11.4‡
KYB	-9.9	-8.8	-8.1	-7.8	-7.7	-7.6	-5.8	-5.4	-5.4	-4.5	-2.3	-1.9	—	1.2	1.2	6.2‡	9.5‡
AIRC	-11.1	-9.9	-9.3	-9.0	-8.9	-8.8	-7.0	-6.6	-6.5	-5.6	-3.5	-3.0	-1.2	—	0.0	5.0‡	8.3‡
ONLINE-G	-11.1	-10.0	-9.3	-9.0	-8.9	-8.8	-7.0	-6.6	-6.6	-5.6	-3.5	-3.1	-1.2	-0.0	—	5.0‡	8.3‡
NLLB_Greedy	-16.2	-15.0	-14.3	-14.1	-13.9	-13.8	-12.0	-11.6	-11.6	-10.7	-8.5	-8.1	-6.2	-5.0	-5.0	—	3.3‡
NLLB_MBR_BLEU	-19.4	-18.3	-17.6	-17.3	-17.2	-17.1	-15.3	-14.9	-14.9	-14.0	-11.8	-11.4	-9.5	-8.3	-8.3	-3.3	—
score	80.7	79.5	78.8	78.6	78.5	78.4	76.6	76.2	76.1	75.2	73.1	72.6	70.8	69.6	69.6	64.5	61.3
rank	1-2	2-6	1-5	2-6	2-5	4-6	7-10	7-10	7-10	7-10	11-12	11-12	13-15	13-15	13-15	16	17

Table 31: Head to head comparison for English→Japanese systems

English→Chinese

	Yishu	Human-refA	GPT4-5shot	Lan-BridgeMT	ONLINE-B	HW-TSC	ONLINE-W	ONLINE-Y	IOL_Research	ONLINE-A	LanguageX	ONLINE-M	ONLINE-G	ANVITA	NLLB_Greedy	NLLB_MBR_BLEU
Yishu	—	0.0	0.1	0.2⋆	0.3	0.7	0.8⋆	2.0⋆	2.3‡	2.5‡	3.6‡	4.0‡	5.0‡	17.7‡	17.9‡	25.0‡
Human-refA	-0.0	—	0.0	0.1†	0.3	0.7	0.8⋆	1.9†	2.3‡	2.5‡	3.6‡	3.9‡	5.0‡	17.7‡	17.8‡	25.0‡
GPT4-5shot	-0.1	-0.0	—	0.1	0.3	0.6	0.7	1.9⋆	2.3‡	2.4‡	3.5‡	3.9‡	5.0‡	17.6‡	17.8‡	24.9‡
Lan-BridgeMT	-0.2	-0.1	-0.1	—	0.2	0.5	0.6	1.8	2.2†	2.3‡	3.4‡	3.8‡	4.9‡	17.5‡	17.7‡	24.8‡
ONLINE-B	-0.3	-0.3	-0.3	-0.2	—	0.3	0.4†	1.6†	2.0‡	2.2‡	3.2‡	3.6‡	4.7‡	17.3‡	17.5‡	24.7‡
HW-TSC	-0.7	-0.7	-0.6	-0.5	-0.3	—	0.1	1.3	1.7‡	1.8‡	2.9‡	3.3‡	4.4‡	17.0‡	17.2‡	24.3‡
ONLINE-W	-0.8	-0.8	-0.7	-0.6	-0.4	-0.1	—	1.2	1.6†	1.7‡	2.8‡	3.2‡	4.3‡	16.9‡	17.1‡	24.2‡
ONLINE-Y	-2.0	-1.9	-1.9	-1.8	-1.6	-1.3	-1.2	—	0.4⋆	0.5†	1.6‡	2.0‡	3.1‡	15.7‡	15.9‡	23.0‡
IOL_Research	-2.3	-2.3	-2.3	-2.2	-2.0	-1.7	-1.6	-0.4	—	0.2	1.2†	1.6†	2.7‡	15.3‡	15.5‡	22.7‡
ONLINE-A	-2.5	-2.5	-2.4	-2.3	-2.2	-1.8	-1.7	-0.5	-0.2	—	1.1⋆	1.5⋆	2.5‡	15.2‡	15.4‡	22.5‡
LanguageX	-3.6	-3.6	-3.5	-3.4	-3.2	-2.9	-2.8	-1.6	-1.2	-1.1	—	0.4	1.5	14.1‡	14.3‡	21.4‡
ONLINE-M	-4.0	-3.9	-3.9	-3.8	-3.6	-3.3	-3.2	-2.0	-1.6	-1.5	-0.4	—	1.1	13.7‡	13.9‡	21.0‡
ONLINE-G	-5.0	-5.0	-5.0	-4.9	-4.7	-4.4	-4.3	-3.1	-2.7	-2.5	-1.5	-1.1	—	12.6‡	12.8‡	20.0‡
ANVITA	-17.7	-17.7	-17.6	-17.5	-17.3	-17.0	-16.9	-15.7	-15.3	-15.2	-14.1	-13.7	-12.6	—	0.2‡	7.3‡
NLLB_Greedy	-17.9	-17.8	-17.8	-17.7	-17.5	-17.2	-17.1	-15.9	-15.5	-15.4	-14.3	-13.9	-12.8	-0.2	—	7.1‡
NLLB_MBR_BLEU	-25.0	-25.0	-24.9	-24.8	-24.7	-24.3	-24.2	-23.0	-22.7	-22.5	-21.4	-21.0	-20.0	-7.3	-7.1	—
score	82.2	82.1	82.1	82.0	81.8	81.5	81.4	80.2	79.8	79.7	78.6	78.2	77.1	64.5	64.3	57.2
rank	1-5	1-5	1-7	3-8	1-6	1-8	4-8	5-8	9-10	9-10	11-13	11-13	11-13	14	15	16

Table 32: Head to head comparison for English→Chinese systems

Japanese→English

	GPT4-5shot	SKIM	Human-refA	ONLINE-Y	ONLINE-B	ONLINE-A	ONLINE-W	NAIST-NICT	GTCOM_DLUT	Lan-BridgeMT	ANVITA	ONLINE-G	LanguageX	ONLINE-M	KYB	AIRC	NLLB_MBR_BLEU	NLLB_Greedy
GPT4-5shot	—	0.7⋆	0.9‡	1.8‡	1.9‡	2.1‡	2.5†	2.9‡	4.4‡	4.8‡	5.5‡	6.5‡	6.7‡	8.4‡	8.9‡	12.4‡	14.6‡	15.2‡
SKIM	-0.7	—	0.2†	1.0⋆	1.2	1.3†	1.7	2.2⋆	3.6‡	4.1‡	4.7‡	5.8‡	5.9‡	7.7‡	8.1‡	11.6‡	13.8‡	14.5‡
Human-refA	-0.9	-0.2	—	0.9	1.0	1.1	1.5	2.0	3.5⋆	3.9‡	4.5‡	5.6‡	5.7‡	7.5‡	7.9‡	11.4‡	13.7‡	14.3‡
ONLINE-Y	-1.8	-1.0	-0.9	—	0.1	0.3	0.7	1.1	2.6†	3.1‡	3.7‡	4.7‡	4.9‡	6.6‡	7.1‡	10.6‡	12.8‡	13.4‡
ONLINE-B	-1.9	-1.2	-1.0	-0.1	—	0.2	0.6	1.0	2.5†	2.9‡	3.6‡	4.6‡	4.8‡	6.5‡	7.0‡	10.5‡	12.7‡	13.3‡
ONLINE-A	-2.1	-1.3	-1.1	-0.3	-0.2	—	0.4	0.8	2.3	2.8‡	3.4‡	4.4‡	4.6‡	6.3‡	6.8‡	10.3‡	12.5‡	13.2‡
ONLINE-W	-2.5	-1.7	-1.5	-0.7	-0.6	-0.4	—	0.4	1.9†	2.4‡	3.0‡	4.0‡	4.2‡	6.0‡	6.4‡	9.9‡	12.1‡	12.8‡
NAIST-NICT	-2.9	-2.2	-2.0	-1.1	-1.0	-0.8	-0.4	—	1.5†	2.0‡	2.6‡	3.6‡	3.8‡	5.5‡	6.0‡	9.5‡	11.7‡	12.3‡
GTCOM_DLUT	-4.4	-3.6	-3.5	-2.6	-2.5	-2.3	-1.9	-1.5	—	0.5‡	1.1†	2.1‡	2.3†	4.0‡	4.5‡	8.0‡	10.2‡	10.9‡
Lan-BridgeMT	-4.8	-4.1	-3.9	-3.1	-2.9	-2.8	-2.4	-2.0	-0.5	—	0.6	1.7	1.8	3.6‡	4.0‡	7.5‡	9.7‡	10.4‡
ANVITA	-5.5	-4.7	-4.5	-3.7	-3.6	-3.4	-3.0	-2.6	-1.1	-0.6	—	1.1	1.2	3.0‡	3.4‡	6.9‡	9.1‡	9.8‡
ONLINE-G	-6.5	-5.8	-5.6	-4.7	-4.6	-4.4	-4.0	-3.6	-2.1	-1.7	-1.1	—	0.2	1.9‡	2.4‡	5.9‡	8.1‡	8.7‡
LanguageX	-6.7	-5.9	-5.7	-4.9	-4.8	-4.6	-4.2	-3.8	-2.3	-1.8	-1.2	-0.2	—	1.8‡	2.2‡	5.7‡	7.9‡	8.6‡
ONLINE-M	-8.4	-7.7	-7.5	-6.6	-6.5	-6.3	-6.0	-5.5	-4.0	-3.6	-3.0	-1.9	-1.8	—	0.5	4.0‡	6.2‡	6.8‡
KYB	-8.9	-8.1	-7.9	-7.1	-7.0	-6.8	-6.4	-6.0	-4.5	-4.0	-3.4	-2.4	-2.2	-0.5	—	3.5‡	5.7‡	6.4‡
AIRC	-12.4	-11.6	-11.4	-10.6	-10.5	-10.3	-9.9	-9.5	-8.0	-7.5	-6.9	-5.9	-5.7	-4.0	-3.5	—	2.2†	2.9†
NLLB_MBR_BLEU	-14.6	-13.8	-13.7	-12.8	-12.7	-12.5	-12.1	-11.7	-10.2	-9.7	-9.1	-8.1	-7.9	-6.2	-5.7	-2.2	—	0.6
NLLB_Greedy	-15.2	-14.5	-14.3	-13.4	-13.3	-13.2	-12.8	-12.3	-10.9	-10.4	-9.8	-8.7	-8.6	-6.8	-6.4	-2.9	-0.6	—
score	81.3	80.6	80.4	79.5	79.4	79.2	78.8	78.4	76.9	76.4	75.8	74.8	74.6	72.9	72.4	68.9	66.7	66.1
rank	1	2-4	3-8	3-8	2-8	3-9	2-8	3-8	8-9	10-13	10-13	10-13	10-13	14-15	14-15	16	17-18	17-18

Table 33: Head to head comparison for Japanese→English systems

Chinese→English

	Lan-BridgeMT	GPT4-5shot	Yishu	ONLINE-W	ONLINE-G	ONLINE-B	ONLINE-Y	HW-TSC	ONLINE-A	IOL_Research	LanguageX	ONLINE-M	NLLB_MBR_BLEU	Human-refA	NLLB_Greedy	ANVITA
Lan-BridgeMT	—	1.9	2.6‡	2.7‡	2.9‡	3.1‡	3.2‡	3.8‡	5.1‡	5.2‡	5.6‡	6.0‡	6.7‡	6.8‡	8.9‡	10.3‡
GPT4-5shot	-1.9	—	0.6‡	0.8‡	1.0‡	1.1†	1.2‡	1.9†	3.1‡	3.3‡	3.7‡	4.1‡	4.7‡	4.9‡	6.9‡	8.3‡
Yishu	-2.6	-0.6	—	0.2	0.3⋆	0.5	0.6	1.3	2.5	2.6‡	3.1‡	3.5‡	4.1‡	4.3‡	6.3‡	7.7‡
ONLINE-W	-2.7	-0.8	-0.2	—	0.2⋆	0.4	0.5	1.1	2.3⋆	2.5‡	2.9‡	3.3‡	4.0‡	4.1‡	6.2‡	7.6‡
ONLINE-G	-2.9	-1.0	-0.3	-0.2	—	0.2	0.3	0.9	2.2	2.3⋆	2.8	3.1‡	3.8‡	3.9‡	6.0‡	7.4‡
ONLINE-B	-3.1	-1.1	-0.5	-0.4	-0.2	—	0.1†	0.8	2.0‡	2.1‡	2.6‡	3.0‡	3.6‡	3.8‡	5.8‡	7.2‡
ONLINE-Y	-3.2	-1.2	-0.6	-0.5	-0.3	-0.1	—	0.7	1.9	2.0†	2.5⋆	2.9‡	3.5‡	3.7‡	5.7‡	7.1‡
HW-TSC	-3.8	-1.9	-1.3	-1.1	-0.9	-0.8	-0.7	—	1.2‡	1.4‡	1.8‡	2.2‡	2.8‡	3.0‡	5.0‡	6.5‡
ONLINE-A	-5.1	-3.1	-2.5	-2.3	-2.2	-2.0	-1.9	-1.2	—	0.1⋆	0.6	1.0‡	1.6‡	1.8‡	3.8‡	5.2‡
IOL_Research	-5.2	-3.3	-2.6	-2.5	-2.3	-2.1	-2.0	-1.4	-0.1	—	0.4	0.8†	1.5‡	1.6‡	3.7‡	5.1‡
LanguageX	-5.6	-3.7	-3.1	-2.9	-2.8	-2.6	-2.5	-1.8	-0.6	-0.4	—	0.4†	1.0‡	1.2‡	3.2‡	4.6‡
ONLINE-M	-6.0	-4.1	-3.5	-3.3	-3.1	-3.0	-2.9	-2.2	-1.0	-0.8	-0.4	—	0.6†	0.8	2.8‡	4.3†
NLLB_MBR_BLEU	-6.7	-4.7	-4.1	-4.0	-3.8	-3.6	-3.5	-2.8	-1.6	-1.5	-1.0	-0.6	—	0.2	2.2	3.6
Human-refA	-6.8	-4.9	-4.3	-4.1	-3.9	-3.8	-3.7	-3.0	-1.8	-1.6	-1.2	-0.8	-0.2	—	2.0⋆	3.4
NLLB_Greedy	-8.9	-6.9	-6.3	-6.2	-6.0	-5.8	-5.7	-5.0	-3.8	-3.7	-3.2	-2.8	-2.2	-2.0	—	1.4
ANVITA	-10.3	-8.3	-7.7	-7.6	-7.4	-7.2	-7.1	-6.5	-5.2	-5.1	-4.6	-4.3	-3.6	-3.4	-1.4	—
score	82.9	80.9	80.3	80.2	80.0	79.8	79.7	79.1	77.8	77.7	77.2	76.9	76.2	76.1	74.0	72.6
rank	1-2	1-2	3-8	3-7	5-10	3-7	4-9	3-8	6-10	10-11	8-11	12-13	13-16	12-15	14-16	13-16

Table 34: Head to head comparison for Chinese→English systems

Findings of the WMT 2023 Biomedical Translation Shared Task: Evaluation of ChatGPT 3.5 as a Comparison System

Mariana Neves[1*] **Antonio Jimeno Yepes**[2] **Aurélie Névéol**[3] **Rachel Bawden**[4]
Giorgio Maria Di Nunzio[11] **Roland Roller**[6] **Philippe Thomas**[6] **Federica Vezzani**[5]
Maika Vicente Navarro[7] **Lana Yeganova**[8] **Dina Wiemann**[9] **Cristian Grozea**[10]
[1]German Centre for the Protection of Laboratory Animals (Bf3R),
German Federal Institute for Risk Assessment (BfR), Berlin, Germany
[2]RMIT University, Australia
[3]Université Paris-Saclay, CNRS, LISN, Orsay, France
[4]Inria, Paris, France
[5]Dept. of Linguistic and Literary Studies University of Padua, Italy
[6]German Research Center for Artificial Intelligence (DFKI), Berlin, Germany
[7]Leica Biosystems, Australia
[8]NCBI/NLM/NIH, Bethesda, USA
[9]Novartis AG, Basel, Switzerland
[10]Fraunhofer Institute FOKUS, Berlin, Germany
[11]Dept. of Information Engineering, University of Padua, Italy

Abstract

We present an overview of the Biomedical Translation Task that was part of the Eighth Conference on Machine Translation (WMT23). The aim of the task was the automatic translation of biomedical abstracts from the PubMed database. It included twelve language directions, namely, French, Spanish, Portuguese, Italian, German, and Russian, from and into English. We received submissions from 18 systems and for all the test sets that we released. Our comparison system was based on ChatGPT 3.5 and performed very well in comparison to many of the submissions.

1 Introduction

We describe the eighth edition of the Biomedical Translation Task[1] that was part of the Eighth Conference on Machine Translation (WMT23). Similar to previous years, we released multiple test sets based on biomedical abstracts that we retrieved from the PubMed database.[2]

We addressed six languages pairs, namely German (de), Spanish (es), French (fr), Italian (it), Russian (ru), and Portuguese (pt), from and into English, as following:

- German into English (de2en) and English into German (en2de);
- Spanish into English (es2en) and English into Spanish (en2es);
- French into English (fr2en) and English into French (en2fr);
- Italian into English (it2en) and English into Italian (en2it);
- Russian into English (ru2en) and English into Russian (en2ru);
- Portuguese into English (pt2en) and English into Portuguese (en2pt).

Different from the previous editions of the shared task, we did not release test sets for Chinese–English or English–Chinese. Novel this year is that we relied on ChatGPT 3.5 to create a performance point of comparison (cf. Section 3), instead of our baseline systems from the previous years.

2 Test sets

We created the test sets following a similar procedure to previous years. We downloaded the set composed of daily update files from Pubmed[3] on

[*]The contribution of the authors are the following: MN prepared the MEDLINE test sets, performed test set validation, manual validation, and organized the shared task; AJY performed test set validation, manual validation, the automatic evaluation and co-organized the shared task; AN compiled information on participants' methods, performed test sets validation, manual validation and annotations of chatGPT outputs on the en2fr test set; RB, GMDN, RR, PT, FV, MVN, LY, DW performed test set validation and/or manual validation; and CG used OpenAI API to create the ChatGPT 3.5 point of comparison; All authors approved the final version of the manuscript. E-mail for contact: mariana.lara-neves@bfr.bund.de

[1]http://www2.statmt.org/wmt23/biomedical-translation-task.html

[2]https://pubmed.ncbi.nlm.nih.gov/

[3]https://ftp.ncbi.nlm.nih.gov/pubmed/updatefiles/

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 43–54
December 6–7, 2023.

April 26, 2023 and searched for articles that contained abstracts in both English and one of the six languages that we consider. We then randomly selected 100 bilingual abstracts for each of the language pairs.

For all language pairs, we split the sentences of the abstracts using SciSpacy (Neumann et al., 2019) and aligned them with the Geometric Mapping and Alignment (GMA) tool.[4] Native speakers of the languages manually checked the alignment quality in the Appraise tool (Federmann, 2018). In this evaluation, we classified the automatically aligned sentences into five categories:

1. "OK": both sentences contain the same information;
2. "Source>Target": the source sentence contains more information than the target one;
3. "Target>Source". the target sentence contains more information than the source one;
4. "Overlap": both source and target sentences have information not contained in the other one;
5. "No Alignment": the sentences refer to completely different contents, or one of hem is missing.

We present the results in Table 1. The highest alignment rates, i.e. the "OK" ones, were for Portuguese (at least 90%, both en2pt and pt2en), and the lowest ones for Russian (only 52% for en2ru). For the latter, we notice that the biggest difference with respect to the other language pairs is that many sentence pairs are not aligned, i.e. the "No Alignment" ones. The percentages for "Source>Target", "Target>Source", and "Overlap" are similar to the other language pairs. An analysis of these errors shows that they are due both to the sentence splitting and the alignment tool.

We released our test sets in two submission systems: (i) our Google form as announced on our shared task's web site; (ii) in OCELoT,[5] both in the General and in the Biomedical test sets.

3 Comparison system - ChatGPT 3.5

Instead of providing a baseline this year, we choose to provide translations from the ChatGPT 3.5 model through the OpenAI API. We refer to ChatGPT as a comparison system rather than a baseline, as it does not satisfy the usual criterion for a baseline as being a transparent, well-understood and reproducible model that provides a good (generally) lower bound against which to compare systems. Notably, the model is closed-sourced and trained on huge amounts of data, of which the details are not openly known.

ChatGPT 4 excels at many tasks (Chen et al., 2023; Jahan et al., 2023), including translation. Researchers from Tencent identified in a limited early evaluation done before the API was available ChatGPT 4 as a good translator (Jiao et al., 2023). Please note that we abstained from using the stronger ChatGPT 4 and used instead the faster but expectedly weaker ChatGPT 3.5. More precisely we used the model snapshot "gpt-3.5-turbo-0613", computed on June 13th 2023 but with the training data "up to Sept 2021"[6]. This reduces the risk of data contamination with respect to the abstracts used in our test sets, which were published in 2023.

The ChatGPT variants are large and trained on large quantities of data, but are generalist systems. Ideally, systems dedicated to translation or specialized in biomedical translation would be able to outperform them, or at least outperform the faster lower-quality version that we proposed here as a point of comparison. Otherwise, there are fewer reasons remaining for developing and using an alternative machine translation (MT) system: data privacy, self-hosting, usage in low-resources, non-connected systems.

We used the following prompt to perform the translations and to keep ChatGPT from producing any comments beyond the translation text itself: **"You are a helpful assistant specialised in biomedical translation. You will be provided with a sentence in {src}, and your task is to translate it into {trg}."** where {src} was the source language and {trg} was the target language (e.g. src = Italian and trg = English).

Using ChatGPT through the API proved to be more challenging than expected and seemed to act as a stress test for the API servers or for the cloudfare content distribution network proxy they use. For example we hit various intentional limitations, such as a rate limit of 90,000 tokens per minute. We then faced multiple other errors: read time out

[4] https://nlp.cs.nyu.edu/GMA/
[5] https://ocelot-wmt23.mteval.org/
[6] https://platform.openai.com/docs/models/gpt-3-5

Language	OK	Source>Target	Target>Source	Overlap	No Align.	Total
de2en	352 (82.2%)	20 (4.7%)	12 (2.8%)	9 (2.1%)	35 (8.2%)	428
en2de	471 (87.7%)	28 (5.2%)	9 (1.7%)	11 (2.0%)	18 (3.4%)	537
es2en	412 (89.5%)	16 (3.5%)	11 (2.4%)	-	21 (4.6%)	460
en2es	388 (88.4%)	21 (4.8%)	15 (3.4%)	6 (1.4%)	9 (2.0%)	439
fr2en	215 (85.3%)	17 (6.7%)	10 (4.0%)	7 (2.8%)	3 (1.2%)	252
en2fr	432 (83.7%)	78 (15.1%)	4 (0.8%)	-	2 (0.4%)	516
it2en	310 (73.4%)	46 (10.9%)	23 (5.5%)	6 (1.4%)	37 (8.8%)	422
en2it	298 (67.0%)	33 (7.4%)	29 (6.5%)	12 (2.7%)	73 (16.4%)	445
pt2en	385 (93.7%)	6 (1.4%)	7 (1.7%)	9 (2.2%)	4 (1.0%)	411
en2pt	450 (90.6%)	21 (4.2%)	12 (2.4%)	9 (1.8%)	5 (1.0%)	497
ru2en	233 (70.0%)	30 (9.0%)	16 (4.8%)	10 (3.0%)	44 (13.2%)	333
en2ru	221 (52.9%)	44 (10.5%)	23 (5.5%)	18 (4.3%)	112 (26.8%)	418

Table 1: Statistics (number of sentences and percentages) of the automatic alignment quality of the MEDLINE test sets.

in the object "HTTPSConnectionPool" with host api.openai.com, HTTP 502 (bad gateway), and "internal error". After writing our API calling code in an idempotent way, we were able to interrupt it whenever it was stuck and restart it whenever we stopped it or it stopped with an error. To this end, the script would skip over the existing translations and proceed with sending for translation, one by one, the rest of the entries not yet translated.

The overall experience remained positive, as building the ChatGPT 3.5 translations involved 674,470 tokens, resulting in a total API cost of only 1.15 USD. However, we have no information on the CO_2 impact of the computation, which should include the impact of inference for translations as well as a fraction of the impact of training the ChatGPT 3.5 model. Writing the scripts and executing them took less than three days. The execution itself was fast; as we reported here, at times we exceeded the API limit of 90,000 tokens per minute.

4 Teams and systems

After the release of the test sets, the teams had around two weeks to process the data and submit their translations. We collected submissions from the two systems (our Google form and OCELoT) belonging to 18 teams (or systems), as listed in Table 2. We allowed up to three runs for each team and language pair. From all submissions, we skipped only one translation from one team, namely the one for fr2en from UPCite-CLILLF, since it was in French (instead of English).

This year, the Google submission form also included questions on material and methods used by participants. The questions were identical to those used in 2022. The response rate was lower than in previous years (2020-2022) when the questionnaire was operated separately from the submission system and teams were asked to complete the survey after submission. In Ocelot submissions, participants were asked to submit a narrative description of their method. None of the teams reported the CO_2 impact of their participation in the task.

Many teams approached the task with transformer-based neural MT (NMT), relying on existing implementations. The use of prompting autoregressive models was also introduced this year. Table 3 presents details of the teams' methods.

5 Automatic evaluation

We present BLEU scores (Papineni et al., 2002) for the automatic evaluation in Tables 4 and 5. This includes translations received from both submission systems (Google Form and OCELoT).

For both en2de and de2en test sets, the submissions from HuaweiTSC, ZengHuiMT, GPT4-5shot, and PROMT teams obtained higher scores than our comparison system (ChatGPT) according to BLEU. The BLEU scores of the Lan-BridgeMT submissions (which use GPT3 and GPT4) came very close to those of ChatGPT for most language pairs, e.g., en2es, en2it, and were sometimes higher, e.g., fr2en, it2en, and ru2en. Most of the ONLINE system submissions also got higher BLEU scores than ChatGPT. However, it is worth bearing in mind the possibility that the ONLINE systems had previously seen our test sets in the large data on

Team ID	Institution	Biom. task	Publication
AIRC	Artificial Intelligence Research Center, Japan	-	(Rikters and Miwa, 2023)
GPT4-5shot	Microsoft	-	(Hendy et al., 2023)
GTCOM_Peter	Global Tone Communication, China	-	(Zong, 2023)
HuaweiTSC	Huawei Translation Service Center	Yes	(Wu et al., 2023)
Lan-BridgeMT	Lan-Bridge Communications, China	Yes	(Wu and Hu, 2023)
NLLB_Greedy	(unknown)	-	-
NLLB_MBR_BLEU	(unknown)	-	-
NRPU_FJWU	Fatima Jinnah Women University, Pakistan	Yes	(Firdous and Rauf, 2023)
ONLINE-A	(unknown)	-	-
ONLINE-B	(unknown)	-	-
ONLINE-G	(unknown)	-	-
ONLINE-M	(unknown)	-	-
ONLINE-W	(unknown)	-	-
ONLINE-Y	(unknown)	-	-
PROMT	PROMT LLC	-	(Molchanov and Kovalenko, 2023)
UPCite-CLILLF	Université Paris Cité, France	Yes	(Zhu et al., 2023)
ustc_ml_group	University of Science and Technology, China	Yes	-
ZengHuiMT	LanguageX, China	-	(Zeng, 2023)

Table 2: List of the participating teams and systems. The third column indicates the teams that directly participated on the Biomedical Translation Task.

Team ID	Language pair	MT method	Trained	Fine-Tuned	BT	LM
AIRC	en/de	Ensemble of Mega transformer models	Yes	No	Yes	Yes
GTCOM	en/de	Transformer model	-	-	-	multilingual models
HuaweiTSC	en/de	Transformer model	-	-	-	-
Lan-BridgeMT	en/de, en/es, en/fr, en/it, en/pt, en/ru	GPT prompting	No	No	No	GPT3, GPT4
NRPU_FJWU	en/fr	Fairseq NMT	No	Yes	No	No
PROMT	en/ru	Marian NMT	Yes	No	-	-
UPCite-CLILLF	en/fr	MBart-50	No	Yes	No	No
USTC	en/fr	Fairseq NMT	Yes	No	No	No
ZengHuiMT	en/de, en/ru	many-to-many encoder decoder transformer model	-	-	-	-

Table 3: Overview of methods used by participating teams. Information is self-reported through the Google/Ocelot submission form for each selected "best run". BT indicates if backtranslation is used and LM if language models were used.

which they were trained, or were used by the authors to assist the production of the abstracts used in the test sets. Although we use the ChatGPT model based on data prior to 2022, meaning that it could not be trained on the parallel abstracts used in the test sets, it is also possible that ChatGPT was used by authors to produce the abstracts that form part of the test set.

6 Manual evaluation

We carried out a manual validation of the quality of the translations for some language pairs using the "3-way ranking" task in the Appraise tool. It consists of a pairwise comparison with three text spans, for example for en2pt: (i) the source text in English, (ii) translation A in Portuguese, and (iii) translation B also in Portuguese. The text is either a sentence or the whole abstract, i.e., we carried out the validation for each sentence and then for the complete abstract.

The evaluator should choose one of the following four options: (i) A=B, i.e., both translations have similar quality; (ii) A>B, i.e., translation A is better than translation B; (iii) A<B, i.e., translation A is worse than translation B; and (iv) error flag in case one or both of the translations do not refer to the same source text.

For the language pairs that we considered, we randomly selected the abstracts until we had at least 100 sentences. We restricted the abstracts

Teams	Runs	en2de	en2es	en2fr	en2it	en2pt	en2ru
AIRC		0.3443					
GPT4-5shot		0.3881					0.3649
HuaweiTSC	run1*	*0.4369					
HuaweiTSC	run2	0.4345					
HuaweiTSC	run3	0.4422					
Lan-BridgeMT		0.3463	0.5098	0.5164	0.4640	0.4832	0.3361
NLLB_Greedy		0.3663					0.3461
NLLB_MBR_BLEU		0.3625					0.3504
ONLINE-A		0.4332					0.4125
ONLINE-B		0.4298					0.4648
ONLINE-G		0.4263					0.3939
ONLINE-M		0.3984					0.3827
ONLINE-W		0.4451					0.4083
ONLINE-Y		0.4075					0.4049
PROMT							0.3872
UPCite-CLILLF				0.2706			
ustc_ml_group	run1			0.4908			
ustc_ml_group	run2*			*0.4998			
ZengHuiMT		0.3883					0.3775
ChatGPT		0.3851	0.5097	0.5318	0.4607	0.5098	0.3513

Table 4: BLEU scores for "OK" aligned test sentences, from English. The submissions without a run number are the ones that were submitted to OCELoT. Primary runs are marked by *.

Teams	Runs	de2en	es2en	fr2en	it2en	pt2en	ru2en
AIRC		0.3714					
GPT4-5shot		0.4371					0.4774
GTCOM_Peter		0.4212					
HuaweiTSC		0.4771					
HuaweiTSC	run1*	*0.4778					
HuaweiTSC	run2	0.4776					
HuaweiTSC	run3	0.4853					
Lan-BridgeMT		0.4215	0.5769	0.4323	0.5272	0.5569	0.4750
NLLB_Greedy		0.4040					0.4386
NLLB_MBR_BLEU		0.3992					0.4437
NRPU_FJWU	run1*			*0.3350			
NRPU_FJWU(1)	run1			0.3082			
NRPU_FJWU	run2			0.2202			
NRPU_FJWU	run3			0.2395			
NRPU_FJWU(1)	run3			0.3350			
ONLINE-A		0.4606					0.5723
ONLINE-B		0.4662					0.4648
ONLINE-G		0.4364					0.5445
ONLINE-M		0.4465					0.4607
ONLINE-W		0.4759					0.4919
ONLINE-Y		0.4075					0.5089
PROMT							0.5156
UPCite-CLILLF							
ustc_ml_group	run1*			*0.4124			
ustc_ml_group	run2			0.3854			
ZengHuiMT		0.4316					0.5256
ChatGPT		0.4360	0.5827	0.4263	0.5067	0.5915	0.4417

Table 5: BLEU scores for "OK" aligned test sentences into English. The submissions without a run number are the ones that were submitted to OCELoT. Primary runs are marked by *.

to those in which the rate of well aligned (OK) sentences was at least 80%. We considered all pairwise combinations from the following translations: (i) the reference translation, as originally available in PubMed, (ii) translations from ChatGPT 3.5, and (iii) translations from systems that directly took part on the Biomedical Translation Task, and not only on the General Task (see Table 2).

We present the results in Tables 6 and 7. We compute a significance test (Wilcoxon test) when

comparing the systems (or reference translation) and we show in bold and with a star (⊛) those cases in which one system (or the reference translation) was better than the other one.

None of the teams could outperform the reference translation for all of the language pairs. Further, for all language pairs that we checked, the quality of the translations from ChatGPT was similar to the reference translation at the sentence level, i.e., there was no significant difference in the results. However, on the abstract level, the ChatGPT translations were found to be better than the reference translations for some language pairs, namely, en2ru and fr2en.

For some of the languages (e.g. en2de), the rankings from the automatic and manual translations appear consistent. The BLEU score from the HuaweiTSC team was much higher than the one from Lan-BrigdeMT (0.43 versus 0.35), and indeed, the quality of the translations from HuaweiTSC was better than the ones from Lan-BrigdeMT. There are however some differences in rankings. For example the manual rankings do not correspond exactly to the automatic rankings for ru2en, fr2en and en2it. Notably, ChatGPT appears to be penalised by BLEU and does better in the manual rankings.

6.1 Quality of the translations

We discuss below, for some language pairs, some of the mistakes that we observed during the manual validation of the submissions.

en2de Similarly to the last few years, the quality of the translations into German was very high. Overall, the individual translations were often similar and differed only in nuances, such as the order of the syntactic constituents. Some models seemed to favour compound nouns more often than others (e.g., Lammellentrennung vs Trennung der Lamellen). However, this usually had no impact on the translation quality. Some systems translated idioms, such as "window of opportunity", literally into German. Especially specialist terms were translated differently by the individual models and it was rather challenging to judge which of the translated terms has better quality (see Example 1).

(1) **en:** The most common surgical fixation options are cerclages and screws, ...
de$_1$: Die häufigsten chirurgischen Fixierungsoptionen sind Zerkel und Schrauben, ...
de$_2$: Die häufigsten chirurgischen Fixierungsmöglichkeiten sind Zuggurte und Schrauben, ...
de$_3$: Die häufigsten operativen Fixationsmöglichkeiten sind Cerclagen und Schrauben, ...

en2es As observed in the last few years, the overall quality of the translations into Spanish was very high. MT systems output was indistinguishable from human translations in many occasions for both systems evaluated: ChatGPT and Lan-Bridge.

The reference translation outperformed Lan-Bridge when evaluating sentences and abstracts. The reference translation was more consistent in the abstracts, had a higher fluency in the translation and a better choice of terminology than Lan-BridgeMT.

For example, "illness recurrence" was translated as "recurrencia' by Lan-Bridge, whereas the reference translation used a more appropriate term "recidiva". Another example in the translation of the term "coronary heart disease", that Lan-Bridge translates literaly as "enfermedad coronaria", while the reference translation uses the medical term "cardiopatía coronaria".

As mentioned, the reference translation was more fluent when compared to Lan-Bridge, oftetimes having a slightly better word order, better concordance subject/verb and using punctuation (commas and full stops) more fluently. Similarly, the reference translation slightly outperformed ChatGPT when comparing abstracts.

However the baseline translation was better than the reference translation at sentence level, this was due to a more overall fluent and consistent translation of abstracts observed in the reference translation when compared to ChatGPT. It must be noted that ChatGPT performed very well compared to the reference translation in most abstracts evaluated manually.

In the following example ChatGPT used the correct punctuation for numbers above 1,000 in Spanish and the reference translation used the incorrect punctuation and was penalized for this fact.

(2) **ChatGPT:** Se incluyeron un total de 22,148 pacientes de 40 estudios.
Reference: Se incluyó un total de 22.148 pacientes de 40 estudios.

When compared against each other, ChatGPT outperformed Lan-Bridge both at the abstract level

Lang. dir.	Pair	Abstracts				Sentences			
		Total	A>B	A=B	A<B	Total	A>B	A=B	A<B
en2de	HuaweiTSC vs. reference	10	0	6	⊛ **4**	100	25	57	17
	HuaweiTSC vs. Lan-BridgeMT	10	⊛ **7**	2	1	100	⊛ **41**	54	5
	HuaweiTSC vs. ChatGPT	10	5	3	2	100	⊛ **29**	59	12
	reference vs. Lan-BridgeMT	10	6	3	1	100	⊛ **32**	54	3
	reference vs. ChatGPT	10	3	6	1	100	18	64	17
	Lan-BridgeMT vs. ChatGPT	10	0	3	⊛ **7**	100	10	61	⊛ **29**
en2es	ChatGPT vs. Lan-BridgeMT	13	4	6	3	107	21	71	15
	ChatGPT vs. reference	13	3	6	4	107	22	65	20
	Lan-BridgeMT vs. reference	13	1	6	6	107	14	69	24
en2fr	reference vs. Lan-BridgeMT	10	⊛ **9**	0	1	108	⊛ **80**	7	21
	reference vs. ChatGPT	10	7	1	2	108	⊛ **71**	5	32
	reference vs. UPCite-CLILLF	10	⊛ **10**	0	0	108	⊛ **107**	0	1
	reference vs. ustc_ml_group	10	⊛ **9**	0	1	108	⊛ **85**	1	21
	Lan-BridgeMT vs. ChatGPT	10	1	5	4	108	24	24	⊛ **60**
	Lan-BridgeMT vs. UPCite-CLILLF	10	⊛ **10**	0	0	108	⊛ **105**	3	0
	Lan-BridgeMT vs. ustc_ml_group	10	⊛ **8**	1	1	108	⊛ **54**	23	31
	ChatGPT vs. UPCite-CLILLF	10	⊛ **10**	0	0	108	⊛ **103**	3	2
	ChatGPT vs. ustc_ml_group	10	⊛ **9**	1	0	108	⊛ **73**	14	20
	UPCite-CLILLF vs. ustc_ml_group	10	0	0	⊛ **10**	108	7	4	⊛ **97**
en2it	Lan-BridgeMT vs. ChatGPT	15	2	1	⊛ **12**	92	16	29	⊛ **47**
	Lan-BridgeMT vs. reference	15	4	1	10	92	25	31	36
	ChatGPT vs. reference	15	9	1	5	92	24	31	37
en2pt	reference vs. Lan-BridgeMT	11	⊛ **5**	6	0	105	35	45	25
	reference vs. ChatGPT	11	4	4	3	105	25	48	32
	Lan-BridgeMT vs. ChatGPT	11	1	5	5	105	18	62	25
en2ru	reference vs. ChatGPT	13	4	3	6	94	8	60	⊛ **25**
	reference vs. Lan-BridgeMT	13	5	4	4	94	22	47	25
	ChatGPT vs. Lan-BridgeMT	13	7	3	3	94	20	64	10

Table 6: Pairwise manual evaluation results for the MEDLINE abstracts test set (from English). We show in bold (and with ⊛) the values which were statistically significant (Wilcoxon test).

Lang. dir.	Pair	Abstracts				Sentences			
		Total	A>B	A=B	A<B	Total	A>B	A=B	A<B
fr2en	NRPU_FJWU vs. reference	19	1	0	⊛ **18**	108	18	6	⊛ **83**
	NRPU_FJWU vs. ustc_ml_group	19	1	1	⊛ **17**	108	19	11	⊛ **78**
	NRPU_FJWU vs. ChatGPT	19	0	0	⊛ **19**	108	3	8	⊛ **97**
	NRPU_FJWU vs. Lan-BridgeMT	19	3	3	⊛ **13**	108	25	11	⊛ **72**
	reference vs. ustc_ml_group	19	⊛ **12**	4	3	108	47	26	34
	reference vs. ChatGPT	19	5	7	7	108	30	26	⊛ **51**
	reference vs. Lan-BridgeMT	19	⊛ **15**	1	3	108	⊛ **60**	19	28
	ustc_ml_group vs. ChatGPT	19	0	1	⊛ **18**	108	13	39	⊛ **56**
	ustc_ml_group vs. Lan-BridgeMT	19	9	3	7	108	45	26	37
	ChatGPT vs. Lan-BridgeMT	19	⊛ **19**	0	0	108	⊛ **69**	30	9
ru2en	ChatGPT vs. reference	13	3	6	4	75	20	41	14
	ChatGPT vs. Lan-BridgeMT	13	⊛ **7**	5	1	75	⊛ **34**	37	4
	reference vs. Lan-BridgeMT	13	⊛ **9**	4	0	75	⊛ **42**	27	6

Table 7: Pairwise manual evaluation results for the MEDLINE abstracts test set (into English). We show in bold (and with ⊛) the values which were statistically significant (Wilcoxon test).

and at the sentence level. As with the reference translation, ChatGPT was more fluent, had a better choice of terminology (domain specific terms) and was more consistent overall at abstract level.

The ChatGPT translation was more fluent in the following example with a better usage of wording. Lan-bridge followed the English source text more closely which made the output less idiomatic.

(3) **ChatGPT:** IO redujo los niveles de glucosa en sangre, restableció el peso corporal y mejoró

la sensibilidad a la insulina, así como la tolerancia a la insulina y a la glucosa en ratones diabéticos.
Lan-bridge: IO redujo los niveles de glucosa en sangre, restableció el peso corporal y mejoró la sensibilidad a la insulina junto con la tolerancia a la insulina y la tolerancia a la glucosa en ratones diabéticos.

While issues are still being observed by the MT systems evaluated manually this year, these are no longer major translation issues as in past years. The issues observed this year for the translations from English to Spanish were minor issues that affect the overall final quality, but can be remediated by editing the MT output to provide better terminology, specially domain specific, more fluent sentences and a better overall consistency in the translation (specially for abstracts).

en2fr Translation quality was somewhat uneven this year. While some translations were very high quality and often similar or identical to reference translations, others exhibited serious issues including inserting erroneous information (see Example 4) or conveying meaning drastically different (see Example 5) or opposite to the original sentence (see Example 6). This type of error can have a severe impact when it results in incorrect medical information (Example 6) or incorrect description of a social group (see Example 5).

(4) **en:** Analysis (. . .) showed that. . .
fr$_1$: L'analyse (. . .) a montré que. . .
fr$_2$: * L'analyse (. . .) a montré que. . . (Traduit par Docteur Serge Messier)

(5) **en:** The criminalization of Black people
fr$_1$: La criminalisation des Noirs
fr$_2$: * La criminalisation des personnes blanches

(6) **en:** blood potassium level ≥ 6.5 mmol/L
fr$_1$: taux de potassium sanguin supérieur à 6,5 mmol/L
fr$_2$: * taux sanguin de potassium inférieur à 6,5 mmol/L

The translation of numerical values was also unreliable: example 5 illustrates the adequate translation of *6.5 mmol/L* into *6,5 mmol/L*, however in another abstract the study population of *52 dogs* was erroneously translated by *54 chiens*.

Issues remain with acronym translation where acronyms are often kept verbatim upon definition (e.g., *developmental disabilities (DD)* translated as *troubles du développement (DD)* instead of the reference translation *troubles du développement (TD)* although consistency seems improved: acronyms, albeit erroneous, are often used throughout a text.

The comparison of translations exhibiting different types of issues also remains difficult. In example 7, although *enquête* is a better translation for *survey* in the context, translation **fr$_1$** was preferred to **fr$_2$** because of the correct translation for *asking about*, which was central to the sentence.

(7) **en:** A survey asking about training
fr$_1$: Un sondage demandant des informations sur la formation
fr$_2$: * Une enquête demandant une formation

Overall, the one-to-one comparisons seemed quite consistent in ranking the systems and reference, and suggest that perhaps the most serious issues identified were concentrated in a few systems.

In addition to the manual evaluation through appraise, a complementary assessment of ChatGPT outputs was conducted, with a focus on *Acronyms* and *Lab Values*, which had been studied in our clinical case descriptions last year. We found that overall, 39 out of 50 test documents contained acronyms and only 3 contained lab values. The low frequency of lab values in the test set suggests that this particular source of translation difficulty for automatic system is not present in random scientifc abstracts. Furthermore, we cannot draw conclusions on the performance of ChatGPT on lab value translations. *Acronym* translations were considered correct when the ChatGPT translation was identical to the reference translation or consisted of an attested acronym use in similar context. Correct acronym translations (74%) included frequent acronyms such as CI (confidence interval), OR (odds ratio) or MRI (magnetic resonance imaging). In other cases, acronyms were either untranslated (16%) or erroneous (10%). These cases included acronyms for terms that were unfrequent or ad-hoc to the documents - albeit often a major topic. It should be noted that they were a source of inconsistent acronym translations in 14 documents - 36% of test documents with acronyms.

fr2en Translation quality was good overall and sometimes indistinguishable from reference translations. Aside from a problem with certain words being dropped at the beginning of translations, sometimes mid-word (quite possibly due to a bug by one or several of the systems), the errors made were similar to previous years.

Term and acronym translation (see Example 8) remained a serious problem and one that was highly influential in reranking decisions, i.e. more so than other errors such as those involving grammar, style or naturalness. In addition to acronym translation errors, we also observed that acronym placement was not always coherent (e.g. an acronym not being defined at the first instance and used consistently afterwards), but in practice this did not influence reranking decisions because of the presence of more serious errors.[7]

(8) **fr:** La migraine est la maladie neurologique la plus fréquemment rencontrée...
en$_1$: Migraine is the most common neurological disorder...
en$_2$: *Mimine is the most frequently encountered neurological disease...

The translation of non-domain-specific terms also posed problem, either those that were ambiguous in context (Example 9), including pronoun translation (for example *sa/son* 'his/her/its/their' being translated as *its* rather than 'his/their' or involving some degree of polarity (Example 10). On a similar note, the omission of words, mainly adjectives and adverbs (e.g. *relativement* 'relatively' and *souvent* 'often') sometimes made the difference between two translations, as did missing final punctuation (when no other errors were present).

(9) **fr:** ...les traitements oraux anciens...
en$_1$: ...older oral treatments...
en$_2$: *...ancient oral treatments...

(10) **fr:** ...un profil d'effets indésirables peu favorable
en$_1$: ...an unfavorable adverse effect profile
en$_2$: *...a slightly favorable side effect profile.

Finally, as in previous years, not all reference translations of were entirely faithful to the French source abstract (paraphrasing, missing or added information). This resulted in some cases in the reference translation being ranked below a system output, including imperfect outputs. Caution should therefore be taken when drawing conclusions about translation quality concerning humans, since intentional paraphrasing by the authors resulted in good abstracts but inferior in terms of our manual evaluation criteria. This partly explains why ChatGPT is "better" than the reference translations for this language pair.

en2it The quality of the translation was on average higher than the previous years. Most of the sentences compared was almost identical and fluent in terms of the quality of language. From a terminological viewpoint, it is possible to identify some inaccuracies in the choice of translating terms in the target language. For example, in *tumour recurrence*, the correct translation of *recurrence* is *recidiva* instead of *ricorrenza*.

Another frequent mistake, which is also a frequent mistake for language learners, is the translation of *hair* in sentences like "hair cortisol concentration (HCC) in healthy and ill cows". In these cases, *hair* must be considered as the hair of animals of body parts, therefore *peli*, and not scalp hair, in Italian *capelli*.

In some cases, there were better choices made by the reference system. For example, in the case of the phrase "[the author] is an initiate into the topic", ChatGPT used *iniziato* to translate *initiate* while a better equivalent would be in this case *novizio* as proposed by the reference system.

Finally, from a syntactic point of view, the results were very similar and only in a few cases we could find a construction that sounded odd or not easy to read. For example, the sentence "Flowmetry data always showed a more or less sudden disappearance of vasomotion." was translated by the reference system with *I dati della flussometria hanno sempre mostrato una più o meno improvvisa scomparsa della vasomotricità* while it would be more appropriate the translation of provided by the baselinte *I dati di flussometria mostravano sempre una scomparsa più o meno improvvisa della vasomozione.*

en2pt The results show that many translations, either from the referenc, ChatGPT, or from the Lan-BridgeMT team, were as good as the reference translation for many sentences (cf. Table 6,

[7] This could be something to look out for in future years when evaluating whole abstracts, when the translation quality allows such fine-grained observations.

"Sentences"). However, there were many cases on which we decide that one passage was better than the other, we discuss some of these differences here.

The most serious mistake that we found was the translation of "back pain" into "pressão arterial" (blood pressure), probably because both of them have the same acronym in English, i.e., "BP".

(11) **en:** The high incidence and worsening of BP ...
pt$_1$: A alta incidência e agravamento do PC ...
pt$_2$: A alta incidência e o agravamento da pressão arterial ...

Similar to previous years, we still found cases in which the English (or simply a wrong) acronym was used (cf. exmple below). Some similar errors might only be noticed when checking the complete text (abstract), and not only single sentences, such as when the translation includes an acronym that was not defined previously.

(12) **en:** ... Creutzfeldt-Jakob disease (CJD) ...
pt$_1$: ... doença de Creutzfeldt-Jakob (DCJ) ...
pt$_2$: ... doença de Creutzfeldt-Jakob (CJD) ...

In some cases, even though both passages were correct, we found that the translation was better due to the use or more medical concepts.

(13) **en:** ... headache attributed to ischemic stroke ...
pt$_1$: A cefaleia atribuída ao acidente vascular cerebral isquêmico ...
pt$_2$: ... a dor de cabeça atribuída ao derrame isquêmico ...

Sometimes the translation included terms that were not suitable, even thought the meaning was close to the source, and it the might have been understood by many readers.

(14) **en:** ... which were analyzed fully and individually.
pt$_1$: ... que foram analisados na íntegra individualmente.
pt$_2$: ... que foram analisados de forma completa e individual.

We chose translation which better describe the facts, depending of the use active or passive voice. Further, in case of passive voice, we preferred caes in which the subjective is closer to the verb, or even before it. We find that it improves the readability.

(15) **en:** The patients underwent magnetic resonance imaging.
pt$_1$: Os pacientes realizaram ressonância magnética. (active voice)
pt$_2$: Os pacientes foram submetidos a ressonância magnética. (passive voice)

(16) **en:** Twelve articles were included in the analysis.
pt$_1$: Foram incluídos na análise 12 artigos.
pt$_2$: Doze artigos foram incluídos na análise.

ru2en While the quality of ru2en translations continues to impress, one recurrent issue centers around the proper handling of abbreviations and acronyms. Often, an acronym is introduced early in the abstract, and holds a clear, defined meaning. Yet, as the text progresses, these acronyms are frequently mishandled by translation systems, failing to link them to their previously established acronym, and frequently transliterating an acronym created in Russian text. This issue manifests itself nearly every time an acronym appears, which makes translations of abstracts that include acronyms not consistently reliable.

For example, the term Ischemic Stroke is introduced in the abstract and abbreviated to "ИИ" which corresponds to the Russian term "ишемического инсульта". One of the reference translations correctly uses the acronym IS to refer to Ischemic Stroke, while the other comes up with an unrelated abbreviation AI.

(17) **ru:** В исследование включили 120 пациентов (57 женщин и 63 мужчины, средний возраст 58,4±6,4 года) в позднем восстановительном периоде ИИ.
en$_1$: The study included 120 patients in the late recovery period of IS, 57 women and 63 men, average age 58.4±6.4 years.
en$_2$: The study included 120 patients (57 women and 63 men, median age 58.4±6.4 years) in the late recovery period of AI.

7 Conclusions

We presented the finding of the edition of the WMT Biomedical Translation Task. We received submission from 18 systems and compared them to translations from ChatGPT 3.5.

In the automatic evaluation, some systems were scored higher than BLEU according to the comparison system (ChatGPT 3.5). In the manual evaluation, none of the systems were systematically better

than the reference translation for all of the language pairs that we evaluated. However, in a couple of cases, namely, for fr2en and en2ru, the translations from ChatGPT were preferred over the reference translations. We presented a details discussion of the errors that we found during the manual evaluation.

Limitations

Our test sets comprise 50 abstracts per language pair/directions. Further, due to the time consuming, difficulty of the task, and number of submissions, the manual evaluation was only carried out for a small sample. However, since our task has been running for eight years, the cumulative number of test sets is satisfactory for testing purposes, and maybe even for few-shot training approaches.

We did not carry out manual evaluation for some of the language pairs (directions), e.g., it2en, for which we do not have experts who are native speakers in the target language and have a very good knowledge in the source language. However, we always release the test sets and the submission files from the participants, with which anyone can carry out further experiments or manual evaluations.

Ethics Statement

Our test sets were derived from PubMed, a database of biomedical citations. These publications are often used in many areas of the medicine, including decision about diagnostic and treatment of patients. Automatic translation in this domain should be used as part of a larger framework that should include human experts for the interpretation of the translations and, if necessary, correct and adapt the text accordingly.

Acknowledgements

Rachel Bawden's participation was funded by her chair position in the PRAIRIE institute funded by the French national agency ANR as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001 and by the Emergence project, DadaNMT, funded by Sorbonne Université.

References

Qijie Chen, Haotong Sun, Haoyang Liu, Yinghui Jiang, Ting Ran, Xurui Jin, Xianglu Xiao, Zhimin Lin, Hongming Chen, and Zhangmin Niu. 2023. An extensive benchmark study on biomedical text generation and mining with ChatGPT. *Bioinformatics*, 39(9):btad557.

Christian Federmann. 2018. Appraise evaluation framework for machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.

Sheema Firdous and Sadaf Abdul Rauf. 2023. Biomedical Parallel Sentence Retrieval using Large Language Models. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How Good are GPT Models at Machine Translation? A comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Huang. 2023. Evaluation of ChatGPT on biomedical tasks: A zero-shot comparison with fine-tuned generative transformers. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 326–336, Toronto, Canada. Association for Computational Linguistics.

Wenxiang Jiao, Wenxuan Wang, JT Huang, Xing Wang, and ZP Tu. 2023. Is ChatGPT a good translator? Yes with GPT-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Alexander Molchanov and Vladislav Kovalenko. 2023. PROMT Systems for WMT23 Shared General Translation Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matīss Rikters and Makoto Miwa. 2023. AIST AIRC Submissions to the WMT23 Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Yangjian Wu and Gang Hu. 2023. Exploring Prompt Engineering with GPT Language Models for Document-Level Machine Translation: Insights and Findings. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Zhanglin Wu, Daimeng Wei, Zongyao Li, Zhengzhe Yu, Shaojun Li, Xiaoyu Chen, Hengchao Shang, Jiaxin Guo, Yuhao Xie, Lizhi Lei, Hao Yang, and Yanfei Jiang. 2023. The Path to Continuous Domain Adaptation Improvements by HW-TSC for the WMT23 Biomedical Translation Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Hui Zeng. 2023. Achieving State-of-the-Art Multilingual Translation Model with Minimal Data and Parameters. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Lichao Zhu, Maria Zimina-Poirot, Maud Bénard, Behnoosh Namdar, Nicolas Ballier, Guillaume Wisniewski, and Jean-Baptiste Yunès. 2023. Training data filtering and fine-tuning strategies - discoveries of UPCite-CLILLF Team's participation in WMT 23 Biomedical Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Hao Zong. 2023. GTCOM Neural Machine Translation Systems for WMT23. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Findings of the WMT 2023 Shared Task on Discourse-Level Literary Translation: A Fresh Orb in the Cosmos of LLMs

Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, Weiyu Chen, Yvette Graham, Bonnie Webber, Philipp Koehn, Andy Way, Yulin Yuan, Shuming Shi
vinnylywang@tencent.com

Abstract

Translating literary works has perennially stood as an elusive dream in machine translation (MT), a journey steeped in intricate challenges. To foster progress in this domain, we hold a new shared task at WMT 2023, the first edition of the *Discourse-Level Literary Translation*. First, we (Tencent AI Lab and China Literature Ltd.) release a copyrighted and document-level Chinese-English web novel corpus. Furthermore, we put forth an industry-endorsed criteria to guide human evaluation process. This year, we totally received 14 submissions from 7 academia and industry teams. We employ both automatic and human evaluations to measure the performance of the submitted systems. The official ranking of the systems is based on the overall human judgments. In addition, our extensive analysis reveals a series of interesting findings on literary and discourse-aware MT. We release data, system outputs, and leaderboard at `http://www2.statmt.org/wmt23/literary-translation-task.html`.

1 Introduction

In past decades, the evolution of machine translation (MT) has undergone significant improvements in accuracy and efficiency, leading to many practical applications in various fields (Bojar et al., 2014; Barrault et al., 2019; Farhad et al., 2021; Kocmi et al., 2022). Despite its success, MT still struggles in certain intricate scenarios to deliver translations that meet high standards (Läubli et al., 2018; Koehn and Knowles, 2017). Translating literary texts is considered to be the greatest challenge for MT due to its complex nature (Toral and Way, 2018; Toral et al., 2018; Ghazvininejad et al., 2018):

- *Rich Linguistic and Cultural Phenomena*: literary texts contain more complex linguistic and cultural knowledge than non-literary ones (Voigt and Jurafsky, 2012; Ghazvininejad et al., 2018). To generate a cohesive and coherent output, MT models require an understanding of the intended meaning and structure of the text at discourse level (Wang et al., 2016, 2018a,b, 2019, 2023b). Furthermore, it demands skillful adaptation of cultural references, idioms, and subtle expressions to capture the essence of the original work in target languages.
- *Limited Data*: existing document-level datasets are news articles and technical documents (Liu and Zhang, 2020; Thai et al., 2022); there is limited availability of copyrighted, discourse-level, parallel data in the literature domain. This makes it difficult to develop models that are able to handle the complexities of literary translation.
- *Long-Range Context*: literature such as novels have much longer contexts than texts in other domains (e.g. news articles). Translation models need to acquire the capacity of modeling long-range context for learning translation consistency and lexical choice (Wang et al., 2017; Wang, 2019; Matusov, 2019; Du et al., 2023).
- *Unreliable Evaluation Methods*: literary evaluation needs to measure the meaning and structure of the text, and the nuances and complexities of the source language. A single automatic evaluation using a single reference is unreliable. Thus, professional translators with well-defined error typologies and targeted automatic evaluation are considered a complement (Matusov, 2019).

With the swift progression of MT and the notable advancements in Large Language Models (LLM) (Ouyang et al., 2022b; OpenAI, 2023), our curiosity is piqued regarding the efficacy of MT and LLM in the realm of literary translation. We aim to explore the extent to which these technologies can aid in addressing the intricate challenges of translating literary works. Therefore, we hold the first edition of the *Discourse-Level Literary Translation* in WMT 2023. Literary texts encompass a wide range of forms, including novels, short stories, poetry, plays, essays, and more. Among

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 55–67
December 6–7, 2023.

Figure 1: The word cloud represents institute and companies from different regions that downloaded the GuoFeng Webnovel Corpus.

these, *web novels*, also known as online or internet novels, represent a unique and rapidly growing subset of literature. Their popularity, accessibility, and diverse genres set them apart. As they provide not only an extensive volume of text but also exhibit distinctive linguistic features, cultural phenomena, and simulations of societies, web novels can serve as valuable resources and challenging for MT research. This year, the shared task mainly focuses on *document-level web novels*, and we introduce a document-level benchmark dataset and establish human evaluation criteria specifically tailored to address the challenges of literary translation:

- **Benchmark Dataset**: We build and release a copyrighted and high-quality Chinese-English training corpus, comprising 2 million sentences sourced from 179 web fictions. This dataset preserves both book-level and chapter-level contexts, and features manually-aligned sentence pairs. We also provide three types of testsets, varying in distribution and document length (in Section 2).
- **Evaluation Methods**: In order to evaluate the translation quality of the participating systems we used both automatic and human evaluation methods. About automatic evaluation, we employ document-level sacreBLEU (d-BLEU) as our metric, which is computed by matching n-grams in the whole document (Liu et al., 2020; Post, 2018). In terms of human evaluation, we propose a well-defined criteria by adapting multidimensional quality metrics (MQM) (Lommel et al., 2014) to fit the context of literary translation. Note that all evaluations are case-sensitive (in Section 3).

Figure 2: Illustration of discourse-level literary translation, which is sampled from our Web Fiction Corpus. Colored words demonstrate rich linguistic phenomena.

We introduce the task overview and submission form in Section 4. This year, 14 submissions were received from 7 different teams, which are detailed in Section 5. We report the evaluation results in Section 6 followed by the conclusion in Section 7.

2 The GuoFeng Webnovel Corpus

We release a copyrighted and high-quality Chinese-English corpus on web novels. Additionally, we provide in-domain pretrained models as supplementary resources. As shown in Figure 1, a total of 45 institutes and companies from various regions have downloaded our dataset, showing that the prposed tasks and data have garnered widespread interest.

2.1 Datasets

Copyright Copyright is a crucial consideration when it comes to releasing literary texts, and it is also one of the primary reasons for limiting the scale of data in this domain. We, Tencent AI Lab and China Literature Ltd., are the copyright owners of the web fictions included in this dataset. In order to promote the advancement of research in this field, we make this data available to the research community, subject to certain terms and conditions.

- After registration, WMT participants can use the corpus for non-commercial research purposes and follow the principle of fair use (CC-BY).

- Modifying or redistributing the dataset is strictly prohibited.
- You should cite the this paper and claim the original download link.

Data Processing The web novels are originally written in Chinese by web novel writers and then translated into English by professional translators. Our data processing involves a combination of automated and manual techniques: 1) we match Chinese books with its English counterparts based on bilingual titles; 2) within each book, Chinese-English chapters are aligned using Chapter ID numbers; 3) within each chapter, we build a MT-based sentence aligner to align sentences in parallel, preserving the sentence order in the chapter; 4) human annotators are engaged to review and correct any discrepancies in sentence-level alignment. To ensure the retention of discourse information, we permit null alignments. We totally spent 6 months addressing copyright issues and around 40,000 euros for human annotation. Figure 2 shows the final format of our corpus.

Training/Validation/Testing Data Table 3 lists data statistics of our dataset. As seen, the *training set* contains 23K continuous chapters from 179 web novels, covering 14 genres such as fantasy science and romance. To enable participants to evaluate model performance by themselves, we provide two *unofficial validation/testing sets* with one reference. For dataset_1, books overlap with the training data, whereas dataset_2 contains unseen books. The participants can regard each chapter as a document to train and test their discourse-aware models. Apart from this, parallel training data in the General MT Task can also be used for data augmentation. In the final testing stage, participants use their systems to translate the *official testing set* (Test_{final}). We select around 20 consecutive chapters from each book. Thus, we participants could treat all chapters within a book as a long document[1]. As seen, the document length of Test_{final} is quite longer than other sets. The final testset contains two references: Reference 1 is translated by human translators and Reference 2 is bult by manually aligning bilingual text in web page. The genres in the valid and test sets are sampled evenly.

[1]The participants can still regard one chapter as a document, which depends on the models' length capability.

2.2 Pretrained Models

Apart from training dataset from web novels, we also provide in-domain pretrained models as supplementary resources. These models can be used to finetune or initialize MT models.

- **RoBERTa (base)**: The original model features a 12-layer encoder and is trained on the Chinese Wikipedia (Liu et al., 2019). It has a hidden size of 768 and a vocabulary size of 21,128 using whole word masking. We continuously train it with Chinese literary texts (84B tokens) (Wang et al., 2023a).
- **mBART (CC25)**: This original model is equipped with a 12-layer encoder and a 12-layer decoder, having been trained on a web corpus spanning 25 languages (Liu et al., 2020). It boasts a hidden size of 1024 and a vocabulary size of 250,000. We continuously train it with English and Chinese literary texts (114B tokens) (Wang et al., 2023a).

Besides, general-domain pretrained models listed in General MT Track are also allowed in this task: mBART, BERT, RoBERTa, sBERT, LaBSE.

3 Evaluation Methods

It is still an open question whether human and automatic evaluation metrics are complementary or mutually exclusive in measuring the document-level and literary translation quality. Thus, we report both automatic and human evaluation methods, and officially rank the systems based on the overall human judgments.

3.1 Automatic Evaluation

We use widely-used sentence- and document-level evaluation metrics: 1) *sentence-level*: we employ sacreBLEU (Post, 2018), chrF (Popović, 2015), TER (Snover et al., 2006) and pretraining-based COMET (Rei et al., 2020); 2) *document-level*: we mainly use document-level sacreBLEU (d-BLEU) (Liu et al., 2020), which is computed by matching n-grams in the whole document. For d-BLEU, We combine all sentences in each document as one line and then conduct sacreBLEU metric. Note that all evaluations are case-sensitive. We employ *sacrebleu*[2] to calculate sacreBLEU, chrF, TER and d-BLEU with *sacrebleu* using two references. The command is: `cat output | python -m`

[2]https://github.com/mjpost/sacrebleu with signature: `nrefs:2|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1`.

Dataset	#Book	#Chap.	#Sent.	#Word	|D|
Train	179	22.6K	1.9M	32.0M	1.4K
Valid_1	22	22	755	18.3K	832
Test_1	26	22	697	19.5K	884
Valid_2	10	10	853	16.0K	1.6K
Test_2	12	12	917	16.7K	1.4K
Test_{final}	12	239	16.7K	337.0K	*28.1K

Figure 3: Data statistics of the GuoFeng Webnovel Corpus on number of book, chapter (#Chap.), sentence (#Sent.), word, and genre distribution in training set. The #Word is based on English texts. For dataset_1, books overlap with the training data, whereas dataset_2 contains unseen books. Thus, each chapter is treated as a separate document. For Test_{final}, around 20 consecutive chapters from each book are selected, treating all chapters within a book as a long document. The document length (|D|) is calculated by dividing #Word divided by the number of documents.

`sacrebleu reference*`. We employ *unbabel-comet*[3] to calculate COMET score using *Reference 1*. The command is: `comet-score -s input -t output -r reference1` (default model).

3.2 Human Evaluation

The human evaluation was performed by professional translators using an adaptation of the multidimensional quality metrics (MQM) framework (Lommel et al., 2014). For example, we consider the preservation of literary style and the overall coherence and cohesiveness of the translated texts. As shown in Table 6, we put forth an industry-endorsed criteria to guide human evaluation process. The main error types are:

- **Accuracy (Acc.)**: The target text does not accurately reflect the source text, allowing for any differences authorized by specifications.
- **Fluency (Flu.)**: Issues related to the form or content of a text, irrespective as to whether it is a translation or not.
- **Style (Sty.)**: The text has stylistic problems.
- **Terminology (Ter.)**: A term (domain-specific word) is translated with a term other than the one expected for the domain or otherwise specified.
- **Locale Convention (Loc.)**: The text does not adhere to locale-specific mechanical conventions and violates requirements for the presentation of content in the target locale.
- **Others (Oth.)**: Other issues such as the signs of MT, gender bias and source errors.

MQM utilizes a scorecard format to quantify the quality assessment results. Evaluators assign numerical values to identified translation errors based on error types, severity, etc., making the assessment results more intuitive. The overall quality score is calculated based on per-word translation accuracy:

$$\text{S} = 1 - \frac{5 \times C_{\text{Min.}} + 10 \times C_{\text{Maj.}} + 25 \times C_{\text{Cri.}}}{\text{Total Word Count}}$$

where where we set four error severity levels: Neutral (Neu.), Minor (Min.), Major (Maj.), Critical (Cri.) with 0/5/10/25 severity penalty. $C_\star$ denotes the number of errors. The "Total Word Count" is calculated based on source input (Chinese word). Considering our task is centered on Zh-to-En translation, we engaged four evaluators who are native English speakers and also fluent in Chinese.

4 Task Description

Overview The shared task will be the translation of literary texts between Chinese→English. Participants will be provided with two types of training datasets: (1) discourse-level GuoFeng Webnovel Corpus; (2) General MT Track Parallel Training Data. Additionally, they are provided two types pretrained models: (1) in-domain pretrained models, including In-domain RoBERTa (base) and In-domain mBART (CC25). (2) other general-domain pretrained models listed in General MT Track. Note that basic linguistic tools are allowed

[3] https://github.com/mjpost/sacrebleu.

in the constrained condition as well as pretrained language models released before February 2023.

In the final testing stage, participants use their systems to translate an official testing set. The translation quality is measured by a manual evaluation and automatic evaluation metrics. All systems will be ranked by human judgement according to our professional guidelines and translators. Participants can submit either constrained (i.e. only use the training data specified above) or unconstrained (i.e. it allows the participation with a system trained without any limitations) systems with flags, and we will distinguish their submissions.

Goals The main goals of the task are to:

- Encourage research in machine translation for literary texts.
- Provide a platform for researchers to evaluate and compare the performance of different machine translation systems on a common dataset.
- Advance the state of the art in machine translation for literary texts.

Submission and Format Submissions will be done by sending us an email to our official email. Each team can submit at most 3 MT outputs per language pair direction, one primary and up to two contrastive. The requirements of submission format are (1) Keep 12 output files that are identical to the testing input files. (2) In the output files, ensure that each line is aligned with the corresponding input line.

5 Participants' and Baseline Systems

Here we briefly introduce each participant's systems and refer the reader to the participant's reports for further details. Table 1 shows the summary of systems and participant teams.

5.1 MaxLab (constrained)

The team from University of Southern California, Information Sciences Institute introduce three translation systems. The *Primary System* is built on a paragraph-level transformer, trained on a paragraph-aligned corpus (with a source side cap of 256 characters), executing translations at the paragraph level. The *Contrastive System 1* deploys a sentence-level transformer, capitalizing on the sentence alignment data available in the datasets. The *Contrastive System 2* adopts a paragraph-level Mega model (Ma et al., 2022). The Mega model proposed a single-head gated attention mechanism equipped with an exponential moving average, which achieves comparable performance compared to Transformers having with fewer parameters. In pre-processing, the team opted for Byte-Pair Encoding (BPE) for tokenization. And they employed Jaccard similarity for sentence alignment during the post-processing phase.

5.2 MAKE-NMT-VIZ (constrained)

The team from Université Grenoble Alpes introduced three translation systems. The *Primary System* finetune the mBART (CC50) model using Train, Valid_1, Test_1 of the GuoFeng Corpus, adopting settings similar to those described by Lee et al. (2022). Specifically, they finetune models for 3 epochs, utilizing the GELU activation function, a learning rate of 0.05, a dropout rate of 0.1, and a batch size of 16. For decoding, a beam search of size 5 was employed. The *Contrastive System 1* is implemented upon a finetuned concatenation transformer (Lupo et al., 2023) with two training steps: (1) a sentence-level transformer is trained for 10 epochs using General, Valid_1, Test_1 datasets; (2) a document-level transformer is finetuned using pseudo-document data (3-sentence concatenation) from Train, Valid_2, Test_2 data for 4 epochs. They use ReLU as an activation function, along with an inverse square root learning rate, a dropout rate of 0.1, and a batch size of 64. For decoding, a beam search of size 4 was employed. The *Contrastive System 2* is a sentence-level transformer model trained for 10 epochs using General, Valid_1, Test_1 datasets. The training adopted an inverse square root scheduled learning rate, a dropout rate of 0.1, and a batch size of 64. Decoding was done using a beam search of size 4.

5.3 TJUNLP (constrained)

The team from Tianjin University introduced a *Primary System* based on a sentence-level Transformer model. The training consists of two phases: initially, it undergoes 100k steps on a dense model, followed by a 50k step fine-tuning on mixture of experts (MOE). They adopt the Polynomial Decay as their learning rate scheduling strategy, with a learning rate set at 2e-4, a dropout rate of 0.1, and a batch size encompassing 4096 tokens. For decoding, a beam search of size 5 was employed. For pre-processing, the team opted for SentencePiece Model (SPM) for tokenization.

ID	Team	Institution	Flag	#System	Main Methods
1	MaxLab	University of Southern California	⊙	3	para-level Transformer
2	MAKE-NMT-VIZ	Université Grenoble Alpes	⊙	3	mBART
3	TJUNLP	Tianjin University	⊙	1	sent-level Transformer
4	DLUT	Dalian University of Technology	⊗	1	GPT-3.5-turbo
5	NTU	Nantong University	⊗	1	Opus-MT
6	HITer-WMT	Harbin Institute of Technology	⊗	2	Llama-7b
7	HW-TSC	Huawei Translation Services Center	⊗	3	doc2doc Transformer

Table 1: The summary of system submission and their participant teams. We also report the number of systems (#System) and the constrained (⊙) and unconstrained (⊗) flags.

5.4 NTU (unconstrained)

The Nantong University team introduce a *Primary System*. It is based on a pretrained MT model, Opus-MT,[4], which is trained on OPUS dataset (Tiedemann and Thottingal, 2020). The model is finetuned on one NVIDIA Tesla A100 80 GB where the learning rate is 5e-5, batch size is 64, max length is 512 and the epoch number is 10.

5.5 DLUT (unconstrained)

The team form Dalian University of Technology introduce a *Primary System* based on GPT-3.5-turbo (Brown et al., 2020). They mainly propose prompt engineering, data filtering, and document segmentation to activate the capabilities of LLMs for discourse-level translation (Zhao et al., 2023).

5.6 HITer-WMT (unconstrained)

The team form Harbin Institute of Technology (Harbin) introduce two translation systems. The *Primary System* centers on instruction fine-tuning, executed through the Llama-7b model within the Parrot framework (Jiao et al., 2023).[5] Specifically, they build an instruction dataset from two comprehensive chapters of our existing training corpus according to methodologies in Peng et al. (2023). This dataset was fine-tuned using Llama-7b over 3 epochs with a learning rate of 2e-5. The *Contrastive System* utilizes the GuoFeng mBART Model provided by the shared task. This model was trained over 10 epochs at a learning rate of 1e-4, with gradient clipping applied to stabilize training.

5.7 HW-TSC (unconstrained)

The team form Huawei Translation Services Center exploit a variety of techniques. They introduce an unconstrained Document-to-Document Translation system. They first train a sentence-level Transformer-big model with a 25-layer encoder and a 6-layer decoder, and perform domain adaptation with novel data on this model. They obtain a strong baseline using data augmentation methods including Back Translation, Forward Translation, and Data Diversification. They then perform incremental training using the Doc2Doc technique to turn the model into a document-level translation model. They also conduct document-level data augmentation using the Multi-resolutional Document-to-Document approach (Sun et al., 2022), and ensue the consistency of NE translations in a document with TrAining Data Augmentation (TADA). They submit three systems: the *Primary System* uses all strategies. In contrast to the primary system, the *Contrastive System 1* system does not use TADA, and the *Contrastive System 2* sets the beam size to 6 during inference, while 10 for other tasks.

5.8 Baseline Systems (unconstrained)

We select three representative systems as baselines. *Commercial Translation System*: we use Google Translate,[6], which usually performs state-of-the-art in translation performance. *Commercial LLM Systems*: we employ GPT-4 (8K) API[7] to translate documents, which is known for its extensive context modeling capabilities (Ouyang et al., 2022a; Wang et al., 2023c). *Open-sourced LLM Models*: we enhance Llama (2K) (Touvron et al., 2023) on document-level translation by using the 200K general-domain document-level training set (Du et al., 2023). All testing were conducted between August 1st and 30th, 2023. In the future, we will use more diverse model architectures such as non-autoregressive translation model (Gu et al., 2017;

[4] https://huggingface.co/Helsinki-NLP/opus-mt-zh-en.

[5] https://github.com/wxjiao/ParroT.

[6] https://translate.google.com.

[7] https://platform.openai.com.

Type	System	Sent-Level				Doc-Level
		BLEU↑	chrF↑	COMET↑	TER↓	d-BLEU↑
Baselines	Llama-MT⋆	n/a	n/a	n/a	n/a	43.1
	GPT-4⋆	n/a	n/a	n/a	n/a	43.7
	Google⋆	37.4	57.0	80.50	57.4	47.3
Primary (con)	MaxLab	34.1	53.3	78.24	62.4	45.0
	MAKE-NMT-VIZ	37.9	56.6	81.50	58.7	**48.0**
	TJUNLP	32.1	51.9	77.93	64.1	43.3
Primary (uncon)	DLUT⋆	40.5	58.5	82.58	54.6	50.2
	NTU⋆	32.3	52.5	78.07	64.3	43.4
	HITer-WMT⋆	16.1	37.1	69.84	80.1	28.0
	HW-TSC⋆	44.3	61.1	82.69	51.8	**52.2**
Contrastive	MaxLab_1	34.5	54.7	79.14	62.7	44.9
	MaxLab_2	33.1	52.4	77.84	63.6	44.4
	MAKE-NMT-VIZ_1	33.8	51.2	76.91	63.5	45.5
	MAKE-NMT-VIZ_2	35.0	52.7	77.26	61.5	46.2
	$\text{HITer-WMT}^{\star}_1$	30.8	49.2	76.41	67.2	40.6
	$\text{HW-TSC}^{\star}_1$	44.6	61.0	82.67	51.8	52.6
	$\text{HW-TSC}^{\star}_2$	44.4	61.5	82.63	52.1	52.2

Table 2: Evaluation results of baseline and participants' systems in terms of **automatic evaluation methods**, including 1) sentence-level metrics BLEU, chrF, COMET, TER; and 2) document-level metrics d-BLEU. Systems marked with ⋆ are unconstrained, while others are constrained. The COMET is calculated with *unbabel-comet* using *Reference 1* while others are calculated with *sacrebleu* using two references. The best primary constrained and unconstrained systems are highlighted.

Type	System	MQM	Rank
Baselines	GPT-4⋆	54.81	1
	Llama-MT⋆	28.40	2
	Google⋆	22.66	3
Primary (con)	MAKE-NMT-VIZ	**42.36**	**1**
	MaxLab	28.58	2
	TJUNLP	18.34	3
Primary (uncon)	DLUT⋆	**63.35**	**1**
	HW-TSC⋆	53.01	2
	NTU⋆	31.66	3
	HITer-WMT⋆	5.56	4

Table 3: Evaluation results of baseline and primary systems in terms of **human evaluation**. We report MQM score and System Rank.

Ding et al., 2020, 2021; Wang et al., 2023d).

6 Evaluation Results

6.1 Automatic Evaluation

We report the automatic evaluation scores of all submissions in Table 2. The evaluation metrics includes 1) sentence-level BLEU, chrF, COMET, TER; and 2) document-level d-BLEU. To calculate d-BLEU, we first concatenate all continuous sentences in one book as on line, and then employ sacreBLEU to obtain scorers. To compute d-BLEU, we merge all the consecutive sentences from a single book into one continuous line, and then utilize the sacreBLEU to generate the scores.

Among constrained Primary systems, the MAKE-NMT-VIZ system shows impressive performance and achieves the best in terms of all metrics. Similarly, the HW-TSC⋆ Primary system achieves the best in constrained settings. As introduced in Section 5, MAKE-NMT-VIZ mainly finetune the mBART pretrained model while HW-TSC⋆ train a doc2doc Transformer model using a number of data augmentation methods.

In the majority of teams, the primary system exhibits superior performance compared to the corresponding contrastive system. The exceptions to this trend are noted in the cases of HITer-WMT⋆ and HW-TSC⋆, where this pattern does not hold. Among the baseline systems, Google Translate, a commercial translation service, outperforms both

Type	Systems	Annotator				Average
		1	2	3	4	
Baselines	GPT-4⋆	95.84	73.38	76.71	87.52	83.36
	Llama-MT⋆	94.18	65.06	78.37	83.36	80.24
	Google⋆	85.02	42.60	59.23	21.13	52.00
Primary (con)	MAKE-NMT-VIZ	97.50	83.36	92.51	91.68	**91.26**
	MaxLab	86.69	61.73	71.71	74.21	73.59
	TJUNLP	88.02	55.07	20.97	69.22	58.32
Primary (uncon)	HW-TSC⋆	91.68	83.36	83.36	91.68	**87.52**
	DLUT⋆	95.01	69.22	84.19	90.02	84.61
	NTU⋆	85.02	39.27	28.45	62.56	53.83
	HITer-WMT⋆	57.57	21.80	0.00	31.78	27.79

Table 4: Analysis of human scores by different annotators on **one sampled document**. We report four annotators' scores and average score of Baselines, primary constrained and unconstrained (⋆) systems.

Annotator	1	2	3	4
1	-	-	-	-
2	0.858	-	-	-
3	0.824	0.878	-	-
4	0.752	0.875	0.676	-
Average	0.902	0.976	0.927	0.891

Table 5: Pearson correlation coefficient between scores by different annotators in Table 4.

commercial and open-source LLMs (GPT-4 API and Llama-MT) in terms of d-BLEU scores. Interestingly, both the top-1 ranked Primary constrained and the top-2 ranked unconstrained systems surpass the performance of the commercial MT system.

6.2 Human Evaluation

Table 3 presents the results of the human evaluation and system rank for the Primary submissions. We enlisted four human annotators to evaluate 5 documents, comprising a total of 2,194 words sourced from distinct books within the final testset for each translation system.

As seen, the MAKE-NMT-VIZ system outperforms the other three constrained systems, while DLUT⋆ ranks first among the four unconstrained systems. This is not fully consistent with the automatic evaluation results in Table 2. Moreover, the top-2 unconstrained systems outperform the best constrained system, highlighting the benefits of external knowledge. This observation is consistent with that of automatic evaluation.

Among the baseline systems, the LLM system performs the best, whereas the MT system shows the poorest performance, diverging from the observations of automatic evaluation. Interestingly, the literary MT-enhanced models perform comparable with some systems such as MaxLab and Google Translate.

6.3 Analysis

Inter-Annotator Agreement We engaged four annotators to independently review an identical document (i.e. 601 words) selected from the testset. Table 4 outlines the individual scores given by each annotator and the corresponding average scores. The findings illustrate that (1) while there is variance in the exact scores assigned by different annotators, their scoring trends align; (2) the results on this sample may diverge from those obtained from a larger dataset, highlighting the necessity of human evaluation on a larger scale.

In our effort to understand the consistency among the human evaluators, we conducted a Pearson correlation analysis on their scoring patterns. Table 5 illustrates the pairwise Pearson correlation coefficients for the scores given by each annotator. The results indicate a high degree of agreement among the annotators. For example, Annotator 2 demonstrated a very high correlation with Annotator 3 ($r = 0.878$) and Annotator 4 ($r = 0.875$). Besides, the Average Scores also reveal strong evaluator consensus on translation quality. This consistency underscores the reliability of the evaluators' judgments across the assessed translations.

Error Type We further analyze the error distribution in human-annotated results. Figure 4 classifies and counts the errors identified in the evaluated

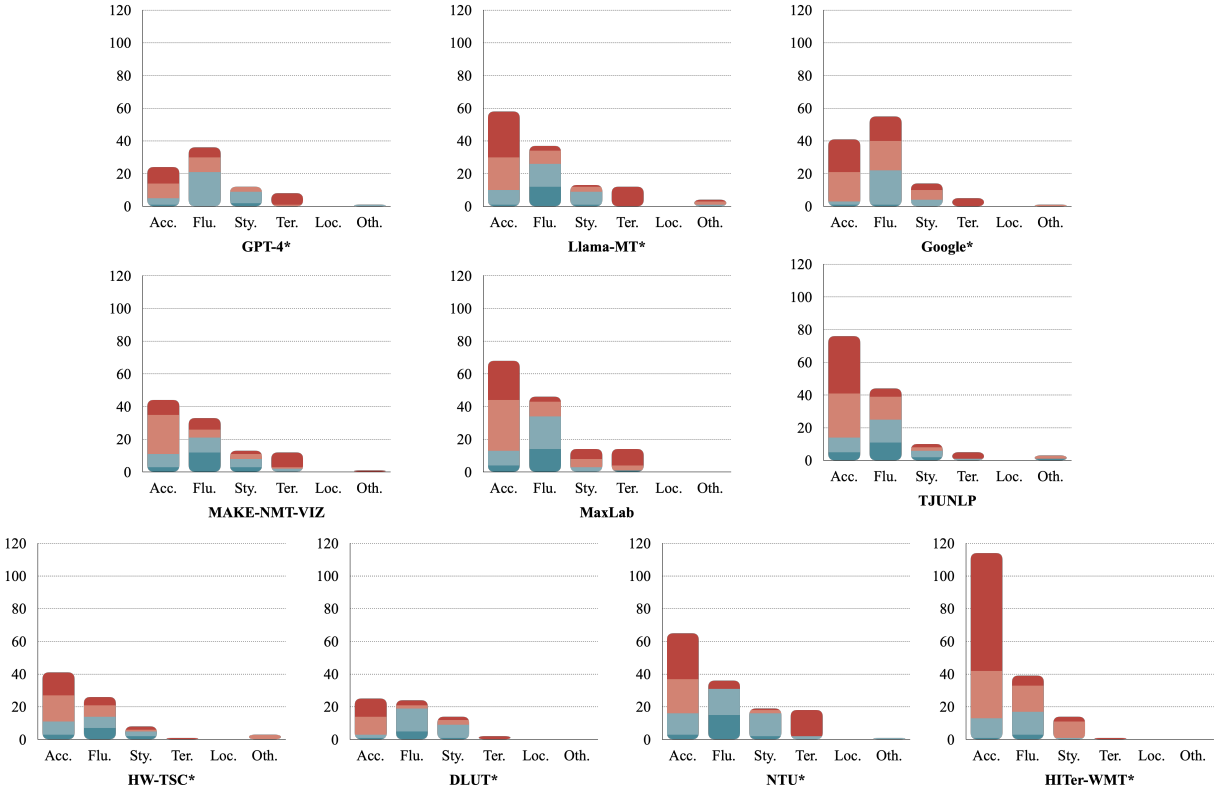


Figure 4: Analysis of error types in human annotations: Accuracy (Acc.), Fluency (Flu.), Style (Sty.), Terminology (Ter.), Localization (Loc.), and Other (Oth.). We report the count of error checkpoints in four evaluated documents. The four error severity levels are presented in different colors: Neutral (blue), Minor (light blue) , Major (light red), Critical (red). Systems marked with ⋆ are unconstrained, while others are constrained.

documents by their severity. This visualization allows for a direct comparison of the error profiles of each system, highlighting their strengths and weaknesses in different aspects of translation quality.

In the baseline systems analysis, GPT-4⋆ registers a higher frequency of Minor errors, particularly in Fluency and Style, indicating areas where refinement could enhance the translation's naturalness and adherence to stylistic norms. Llama-MT⋆, by contrast, has a pronounced incidence of Major and Critical errors in Accuracy and Terminology, raising concerns about the fidelity and technical precision of its translations. Google⋆ stands out with its Fluency errors, suggesting potential issues in maintaining a coherent and natural flow compared to the language models.

Regarding the constrained systems, MAKE-NMT-VIZ displays an even spread of errors, with relatively fewer instances in each category, which points to a well-rounded performance in capturing nuances across various aspects of translation. Both MaxLab and TJUNLP exhibit an increased number of Accuracy and Fluency errors, suggesting challenges in delivering translations that are not only faithful to the source material but also exhibit a seamless and natural flow in the target language.

The unconstrained systems, particularly HW-TSC⋆ and DLUT⋆, show a notable reduction in errors related to Accuracy and Fluency when compared to their constrained counterparts. This trend suggests that the lack of constraints may afford these systems more flexibility, resulting in translations that are more accurate and fluid. However, the overall error distribution across different systems highlights the complex trade-offs and challenges inherent in machine translation, underscoring the need for continued innovation and optimization in the field. In the future, we will also consider hallucination errors (Zhang et al., 2023).

7 Conclusion and Future Work

We believe that the WMT2023 Shared Task on discourse-level literary translation will be a valuable contribution to the field of machine translation and will encourage further research in this area. We discuss the potential limitations of this edition of the shared task as follows:

- *Language Pair*. This year, we only focus on Chinese→English direction. However, we have a long-term plan to continuously organize this task, and will extend the copyrighted dataset into Chinese-Russian and Chinese-German language pairs next year.
- *Literary Genre*. This year, we mainly used the Web Fiction Corpus which is only one type of literary text. We use Web Fiction for two reasons: (1) its literariness is less complicated than others (e.g. poetry, masterpiece); (2) such bilingual data are numerous and continuously increased. We will consider to extend more literary genres such as poetric translation in the next year.
- *Discourse Benchmark*. We have accumulated some discourse- and context-aware benchmarks (Xu et al., 2022, 2023; Wang et al., 2023a). These benchmarks are pivotal for assessing the proficiency of LLMs in handling complex language structures and contextual nuances. As participation of LLM-based systems in our shared tasks increases, we anticipate integrating these benchmarks more comprehensively into our future evaluations to better measure and understand the evolution of LLM capabilities in linguistic context and discourse comprehension.

Machine translation of web novels not only holds research value but also offers practical application prospects (Huang et al., 2021; Lyu et al., 2023). This shared task serves to spur competitive innovation and fosters the advancement of sophisticated machine translation systems capable of navigating the intricate nuances of literary works. Anticipating the future, our objective is to broaden the engagement in the forthcoming shared task, inviting an extensive range of collaborators from industry and academia alike to contribute their unique insights and expertise.

Acknowledgements

We would like to thank the WMT2023 organizers for providing us the opportunity to explore this new task. We also express our gratitude to the experts on the Shared Task Committee for their efforts in organization, evaluation, and advisory roles:

- Longyue Wang, Zhaopeng Tu, Dian Yu, Chenyang Lyu, Shuming Shi (Tencent AI Lab)
- Yan Gu, Yufeng Ma, Weiyu Chen (China Literature Ltd.)
- Bonnie Webber (University of Edinburgh)
- Siyou Liu, Yulin Yuan (University of Macau)
- Philipp Koehn (Johns Hopkins University)
- Liting Zhou, Andy Way (Dublin City University)
- Yvette Graham (Trinity College Dublin)
- Chao-Hong Liu (Potamu Research Ltd.)
- Qingsong Ma (Tencent AI Evaluation Lab)

References

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation*.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Liang Ding, Longyue Wang, Xuebo Liu, Derek F Wong, Dacheng Tao, and Zhaopeng Tu. 2020. Understanding and improving lexical choice in non-autoregressive translation. *arXiv preprint arXiv:2012.14583*.

Liang Ding, Longyue Wang, Xuebo Liu, Derek F Wong, Dacheng Tao, and Zhaopeng Tu. 2021. Progressive multi-granularity training for non-autoregressive translation. *arXiv preprint arXiv:2106.05546*.

Zefeng Du, Wenxiang Jiao, Longyue Wang, Chenyang Lyu, Jianhui Pang, Leyang Cui, Kaiqiang Song, Derek F. Wong, Shuming Shi, and Zhaopeng Tu. 2023. On extrapolation of long-text translation with large language models. *arXiv preprint*.

Akhbardeh Farhad, Arkhangorodsky Arkady, Biesialska Magdalena, Bojar Ondřej, Chatterjee Rajen, Chaudhary Vishrav, Marta R Costa-jussa, España-Bonet Cristina, Fan Angela, Federmann Christian, et al. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*.

Marjan Ghazvininejad, Yejin Choi, and Kevin Knight. 2018. Neural poetry translation. In *NAACL*.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. 2017. Non-autoregressive neural machine translation. *arXiv preprint arXiv:1711.02281*.

Guoping Huang, Lemao Liu, Xing Wang, Longyue Wang, Huayang Li, Zhaopeng Tu, Chengyan Huang, and Shuming Shi. 2021. Transmart: A practical interactive machine translation system. *arXiv preprint arXiv:2105.13072*.

Wenxiang Jiao, Jen tse Huang, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Parrot: Translating during chat using large language models tuned with human translation and feedback. In *Findings of EMNLP*.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, et al. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation*.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *EMNLP*.

En-Shiun Annie Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Ifeoluwa Adelani, Ruisi Su, and Arya D McCarthy. 2022. Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation? *arXiv preprint arXiv:2203.08850*.

Siyou Liu and Xiaojun Zhang. 2020. Corpora for document-level neural machine translation. In *LREC*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463.

Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. 2023. Encoding sentence position in context-aware neural machine translation with concatenation. In *The Fourth Workshop on Insights from Negative Results in NLP*.

Chenyang Lyu, Jitao Xu, and Longyue Wang. 2023. New trends in machine translation using large language models: Case examples with chatgpt. *arXiv preprint arXiv:2305.01181*.

Xuezhe Ma, Chunting Zhou, Xiang Kong, Junxian He, Liangke Gui, Graham Neubig, Jonathan May, and Luke Zettlemoyer. 2022. Mega: moving average equipped gated attention. *arXiv preprint arXiv:2209.10655*.

Evgeny Matusov. 2019. The challenges of using neural machine translation for literature. In *Proceedings of the qualities of literary machine translation*, pages 10–19.

OpenAI. 2023. GPT-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, et al. 2022a. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022b. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *EMNLP*.

Mathew Snover, Bonnie Dorr, Richard Schwartz, John Makhoul, Linnea Micciulla, and Ralph Weischedel. 2006. A study of translation error rate with targeted human annotation. In *Proceedings of the Association for Machine Transaltion in the Americas (AMTA)*.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. Rethinking document-level neural machine translation. In *Findings of ACL*.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. *arXiv preprint arXiv:2210.14250*.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*.

Antonio Toral and Andy Way. 2018. What level of quality can neural machine translation attain on literary text? In *Translation Quality Assessment*, pages 263–287. Springer.

Antonio Toral, Martijn Wieling, and Andy Way. 2018. Post-editing effort of a novel with statistical and neural machine translation. *Frontiers in Digital Humanities*, 5:9.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Rob Voigt and Dan Jurafsky. 2012. Towards a literary machine translation: The role of referential cohesion. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*.

Longyue Wang. 2019. *Discourse-aware neural machine translation*. Ph.D. thesis, Ph. D. thesis, Dublin City University, Dublin, Ireland.

Longyue Wang, Zefeng Du, Donghuai Liu, Cai Deng, Dian Yu, Haiyun Jiang, Yan Wang, Leyang Cui, Shuming Shi, and Zhaopeng Tu. 2023a. Disco-bench: A discourse-aware evaluation benchmark for language modelling. *arXiv preprint arXiv:2307.08074*.

Longyue Wang, Siyou Liu, Mingzhou Xu, Linfeng Song, Shuming Shi, and Zhaopeng Tu. 2023b. A survey on zero pronoun translation. *arXiv preprint arXiv:2305.10196*.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023c. Document-level machine translation with large language models. *arXiv preprint arXiv:2304.02210*.

Longyue Wang, Zhaopeng Tu, Shuming Shi, Tong Zhang, Yvette Graham, and Qun Liu. 2018a. Translating pro-drop languages with reconstruction models. In *AAAI*.

Longyue Wang, Zhaopeng Tu, Xing Wang, and Shuming Shi. 2019. One model to learn both: Zero pronoun prediction and translation. In *EMNLP-IJCNLP*.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. In *EMNLP*.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2018b. Learning to jointly translate and predict dropped pronouns with a shared reconstruction mechanism. In *EMNLP*.

Longyue Wang, Zhaopeng Tu, Xiaojun Zhang, Hang Li, Andy Way, and Qun Liu. 2016. A novel approach to dropped pronoun translation. In *NAACL*.

Zhihao Wang, Longyue Wang, Jinsong Su, Junfeng Yao, and Zhaopeng Tu. 2023d. Revisiting non-autoregressive translation at scale. *arXiv preprint arXiv:2305.16155*.

Mingzhou Xu, Longyue Wang, Siyou Liu, Derek F Wong, Shuming Shi, and Zhaopeng Tu. 2023. A benchmark dataset and evaluation methodology for chinese zero pronoun translation. *Language Resources and Evaluation*.

Mingzhou Xu, Longyue Wang, Derek F. Wong, Hongye Liu, Linfeng Song, Lidia S. Chao, Shuming Shi, and Zhaopeng Tu. 2022. GuoFeng: A benchmark for zero pronoun recovery and translation. In *EMNLP*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Anqi Zhao, Kaiyu Huang, Hao Yu, and Degen Huang. 2023. Dutnlp system for wmt23 discourse-level literary translation. In *Proceedings of the Eighth Conference on Machine Translation*.

Type	Granular	Definition	Examples
Accuracy	Addition	The target text includes text not present in the raw.	A translation includes portions of another translation that were inadvertently pasted into the document or the translator has added too many details of his own.
	Omission	Content is missing from the translation that is present in the source.	A paragraph present in the source is missing in the translation.
	Mistranslation	The target content does not accurately match the raw.	A source text states that a medicine *should not be* administered in doses greater than 200 mg, but the translation states that it should be administered in doses greater than 200 mg (i.e., negation has been omitted).
	Misnomer	The target text is more/less specific than the raw.	1. The source text refers to a boy but is translated with a word that applies only to young boys rather than the more general term. 2. The source text uses words that refer to a specific type of military officer but the target text refers to military officers in general.
	Untranslated	Content that should have been translated has been left untranslated.	A sentence to be translated into English was left in Chinese.
Fluency	Punctuation	Punctuation marks missing or used in a wrong way.	An English text uses a semicolon where a comma should be used.
	Spelling	Issues related to spelling of words. (Including those of capitalization, hyphenated words, and use of asterisk for censored swear words.)	The English word "Translation" is spelled "Transaltion".
	Grammar	Issues related to the grammar or syntax of the text, other than spelling and orthography. (especially inconsistency of the tenses and conditionals)	An English text reads "The man was seeing the his wife."
	Inconsistency	The text shows internal inconsistency.	A text uses both "app." and "approx." for "approximately".
Style	Awkwardness	A text is written with an awkward style.	A text is written with many embedded clauses and an excessively wordy style. While the meaning can be understood, the text is very awkward and difficult to follow.
	Inconsistent	Style is inconsistent within a text.	One part of a text is written in a light and terse style while other sections are written in a more wordy style.
	Unidiomatic	The content is grammatical, but not idiomatic.	The following text appears in an English translation of "我们衷心感谢他": "We thanked him with heart" where "with heart" is an understandable, but non-idiomatic rendering, better stated as "heartily".
Terminology	Mistranslation	A genre-specific or cultural-specific terminology is wrongly translated.	A Chinese word "修士" is translated into "practitioner" rather than the expected "cultivator".
	Inconsistent	Terminology is used in an inconsistent manner within the text.	"斗罗大陆" is translated into "Douluo Land" in the first few chapters and then into "Soul Land".
Locale Convention	Location Format	Using the wrong format for address, name etc.	A Chinese address "北京市朝阳区花园路22号" is translated into "Beijing, Chaoyang district, Huayuan Road N.22" instead of the expected "N.22, Huayuan Road, Chaoyang District, Beijing".
	Number Format	The translated date, time, currency, telephone use formats inappropriate for its locale.	An English text has 2012-06-07 instead of the expected 06/07/2012.
Others		Other issues that haven't been included in this list.	E.g. signs of MT, mimetic word, gender bias, source errors etc.

Table 6: The MQM-based evaluation criteria for literary translation.

Findings of the Second WMT Shared Task on Sign Language Translation (WMT-SLT23)

Mathias Müller
University of Zurich

Malihe Alikhani
Northeastern University

Eleftherios Avramidis
DFKI Berlin

Richard Bowden
University of Surrey

Annelies Braffort
University of Paris-Saclay

Necati Cihan Camgöz
Meta Reality Labs

Sarah Ebling
University of Zurich

Cristina España-Bonet
DFKI Saarbrücken

Anne Göhring
University of Zurich

Roman Grundkiewicz
Microsoft

Mert Inan
Northeastern University

Zifan Jiang
University of Zurich

Oscar Koller
Microsoft

Amit Moryossef
Bar-Ilan University

Annette Rios
University of Zurich

Dimitar Shterionov
Tilburg University

Sandra Sidler-Miserez
HfH Zurich

Katja Tissi
HfH Zurich

Davy Van Landuyt
European Union of the Deaf

Abstract

This paper presents the results of the Second WMT Shared Task on Sign Language Translation (WMT-SLT23)[1]. This shared task is concerned with automatic translation between signed and spoken[2] languages. The task is unusual in the sense that it requires processing visual information (such as video frames or human pose estimation) beyond the well-known paradigm of text-to-text machine translation (MT). The task offers four tracks involving the following languages: Swiss German Sign Language (DSGS), French Sign Language of Switzerland (LSF-CH), Italian Sign Language of Switzerland (LIS-CH), German, French and Italian. Four teams (including one working on a baseline submission) participated in this second edition of the task, all submitting to the DSGS-to-German track. Besides a system ranking and system papers describing state-of-the-art techniques, this shared task makes the following scientific contributions: novel corpora and reproducible baseline systems. Finally, the task also resulted in publicly available sets of system outputs and more human evaluation scores for sign language translation.

1 Introduction

This paper presents the outcome of the Second WMT Shared Task on Sign Language Translation (WMT-SLT23). This shared task focuses on automatic translation between signed and spoken languages. Our main goal is working towards including signed languages in NLP research (Yin et al., 2021).

Sign language translation requires processing visual information (such as video frames or human pose estimation) beyond the well-known paradigm of text-to-text machine translation (MT). As a consequence, viable solutions need to consider a combination of Natural Language Processing (NLP), computer vision (CV), computer graphics and animation techniques.

We build on and extend the work done for the first shared task on sign language translation (WMT-SLT22; Müller et al., 2022). Compared to the first edition, we

- extended our competition to more languages (three language pairs instead of one),
- provided much more training data for Swiss German Sign language compared to last year (437 hours instead of 16),
- emphasized sign languages as the target language instead of the source, for instance, by offering official baseline systems for spoken-to-signed translation (not offered last year).

In this second edition of the shared task, we considered the following languages: Swiss German Sign Language (DSGS), French Sign Language of Switzerland (LSF-CH), Italian Sign Language

[1] https://www.wmt-slt.com/

[2] In this paper we use the word "spoken" to refer to any language that is not signed, no matter whether it is represented as text or audio, and no matter whether the discourse is formal (e.g. writing) or informal (e.g. dialogue).

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 68–94
December 6–7, 2023.

of Switzerland (LIS-CH), German, French, and Italian. We offered four tracks: DSGS-to-German translation, German-to-DSGS translation, French-to-LSF translation, and Italian-to-LIS translation.

Four teams participated in the task, which we consider a success. All teams submitted to the DSGS-to-German track, while there were no submissions to any of the tracks where a sign language is the target language.

The remainder of this paper is organized as follows:

- We give some background on sign languages and sign language processing in §2.
- We describe the shared task tracks and submission procedure in §3.
- We report on the corpora we built and distributed specifically for this task in §4 and §5.
- We describe all submitted systems, including our baselines in §6.
- We ran both an automatic and a human evaluation. We explain our evaluation in §7.
- We share the main outcomes in §8 and discuss in §9.

2 Background

In recent years, Sign Language Processing (SLP) has emerged as a sub-area of Natural Language Processing (NLP). Within this field, automatic sign language translation (SLT; or sign language machine translation, SLMT) represents a more specialized discipline, aiming to develop technology that facilitates translation between sign languages and spoken or written languages, but also between sign and sign languages. However, the challenges related to SLP and SLT differ from those of NLP and MT for spoken languages in both range and complexity. Due to the different modality, lack of structured, high-quality, high-quantity data, and the lack of NLP tools, joint efforts from the fields of sign linguistics and computational linguistics, computer science, machine learning, computer vision, 3D animation and others are needed in order to advance this field.

In this section we give an introduction to sign languages (§2.1) and describe the societal and academic relevance of SLP (§2.2). Then we give an overview of SLP in general (§2.3) and of SLT in particular (§2.4) For a general motivation for a shared task involving sign languages see Müller et al. (2022).

2.1 Sign languages

Sign languages are natural languages with their own grammatical structures and lexicons, primarily used by the deaf and hard-of-hearing communities. Contrary to the popular belief that sign language is universal, hundreds of different SLs have been documented so far.

Nature of sign languages Sign languages are visuo-gestural languages. A signer conveys an utterance using their body: through the expression of manual features (hand configuration, location, and orientation) and non-manual features (including facial expressions, mouthing and mouth gestures, gaze and torso direction). The linguistic system of SLs makes use of these specific channels. Information is expressed simultaneously (as opposed to the sequential nature of spoken language), organized in three-dimensional space, and iconicity plays a central role (Woll, 2013; Perniss et al., 2015; Slonimska et al., 2021).

Writing systems To date, SLs have no universally accepted written form or graphical system for transcription (Pizzuto and Pietrandrea, 2001; Filhol, 2020). Several notation systems, such as HamNoSys (Hanke, 2004) or SignWriting (Sutton, 1990; Bianchini and Borgia, 2012), are used in research or teaching but are rarely adopted as a writing system in everyday life, limiting the standardisation of data collection and processing. In SL research, a common practice is therefore to use glosses – text-based, semantic labels for signs, typically borrowed from the corresponding regional spoken language.

A common misconception among MT researchers is that transcribed glosses are a full-fledged writing system for sign languages. In reality, glossing can only be seen a linguistic tool, useful for annotating corpora for linguistic studies (Johnston, 2010). Glosses do not adequately represent the meaning of an SL utterance and, more importantly, "deaf people do not read or write glosses" in everyday life (Müller et al., 2023).

2.2 Relevance of sign language processing

SLP is a research area with high potential societal and academic impact.

Societal impact The overall aim of SLP is to provide language technology for sign languages, which currently are somewhat overlooked, since the vast majority of NLP systems are designed only for spoken languages. This means that more research in SLP could result in more equal access to language technology.

The more specific goal of SLT is to facilitate communication between the deaf and hard-of-hearing communities on the one side and the hearing community on the other side. There is a need for this because speakers of spoken languages and signers of sign languages experience communication difficulties (the same kind of difficulties encountered by speakers of different spoken languages). We emphasize that these technologies should be developed in such a way, so that deaf/hard-of-hearing and hearing people can benefit from them in an equal measure.[3]

Besides aiding direct communication, SLT would improve accessibility to spoken language content, given that spoken languages are often a second language for deaf people, where they exhibit varying proficiency. The reverse direction is also crucial, for example to automatically subtitle signed content to make it accessible to people who do not know sign languages (Bragg et al., 2019).

Academic relevance In the field of NLP, working on sign languages is highly innovative and timely. Recently, a call for more inclusion of signed languages in NLP (Yin et al., 2021) was widely publicized, and an ACL initiative for Diversity and Inclusion[4] targets SL processing as well.

And even though sign languages are still a niche topic in the general field of NLP (the vast majority of NLP systems are designed for spoken languages, not for signed languages), the advancement and spread of SLP tools, calls, initiatives and events lead to knowledge transfer not only within the academic spheres, or between researchers, developers and users, but also, more importantly, between deaf, hard-of-hearing and hearing individuals involved in the process.

[3]We distance ourselves from the audistic view that only deaf people are in need (of access to spoken language discourse). Language barriers are inherently two-way, and addressing them involves both parties.

[4]https://www.2022.aclweb.org/dispecialinitiative

2.3 Sign language processing

Sign language processing is an interdisciplinary field, bringing together research on NLP and computer vision, among other disciplines (Bragg et al., 2019). For a general overview in the context of NLP see Yin et al. (2021); Moryossef and Goldberg (2021).

Tasks SLP involves a variety of (sub)tasks with individual challenges. Widely known tasks are sign language recognition, sign language translation, and sign language production (or *synthesis*). Sign language recognition usually refers to identifying individual signs from videos; see Koller (2020) for an overview. Sign language translation refers to the task of transforming sign language data to a second language, no matter whether signed or spoken; see De Coster et al. (2022) for a comprehensive survey. Finally, sign language production refers to rendering sign language as a video, using methods such as avatar animation (Wolfe et al., 2022) or video generation.

SLP research is challenging for a number of different reasons. The ones we chose to highlight here are linguistic properties, availability of data, and availability of basic NLP tools.

Linguistic challenges SLP is challenging because the characteristics of sign languages (§2.1) cannot be fully handled with existing methods, for instance, the multilinearity, the use of the signing space, and the iconicity. As explained earlier, SLP needs to take into account manual and non-manual cues in order to capture a complete linguistic picture of an SL utterance (Crasborn, 2006). Information is spatio-temporal in nature and the data is simultaneously conveyed by a number of articulators. Signing makes frequent use of indexing strategies for example to identify referents introduced earlier in the discourse or timelines (Engberg-Pedersen, 1993). In other words, a sign language utterance is not a simple sequence of lexical units.

Sign languages have an established vocabulary but are also lexically productive to allow for the definition of new signs or constructions to be used to depict entities or situations (Johnston, 2011).

Availability of data Given the current research landscape in NLP, sign languages are under-resourced. An analysis by Joshi et al. (2020) places all sign languages considered in this study in the category "left behind" (together with many spoken

languages). Existing resources are small and heterogeneous. They are created under a variety of circumstances and vary in quality (e.g. video resolution), signer demographics (e.g. deaf vs. hearing signers), richness of annotation (e.g. glosses, sentence segmentation, translation to a spoken language), and linguistic domain (e.g. only weather reports, hence a very limited domain).

Also, not all corpora are easily accessible online and some have restrictive licenses that disallow NLP research. A survey of SL corpora available in Europe can be found in Kopf et al. (2021). For an account of further challenges relating to data see De Sisto et al. (2022).

Lack of basic linguistic tools SLP currently lacks fundamental NLP tools that are readily available for spoken languages. Such tools include automatic language identification (Monteiro et al., 2016), sign segmentation (De Sisto et al., 2021), sentence segmentation (Ormel and Crasborn, 2012; Bull et al., 2020b) and sentence alignment (Varol et al., 2021). Although there are experimental solutions, they are not yet viable in practice.

Tools like these would be crucial to create better corpora by constructing them automatically, as is routinely done for spoken languages (Bañón et al., 2020), and develop better high-level NLP solutions.

2.4 Sign language translation

In recent years, different methods to tackle SLT have been proposed, most of them suggesting a cascaded system where a signed video is first converted to an intermediate representation and then to spoken text (similarly for text-to-video translation). Intermediate representations (with individual strengths and weaknesses) include pose estimation (§5.3), glosses or writing systems such as HamNoSys (§2.1, writing systems).

There is existing work on gloss-to-text translation (e.g. Camgöz et al. 2018; Yin and Read 2020) and vice versa (e.g. Stoll et al., 2020), pose-to-text translation and vice versa (e.g. Ko et al. 2019; Saunders et al. 2020a,b,c; Inan et al. 2022; Viegas et al. 2023) and systems involving HamNoSys (e.g. Morrissey 2011; Walsh et al. 2022), or AZee expressions, designed to be used as input to avatar synthesis systems (Bertin-Lemée et al., 2023). Recently, direct video-to-text translation was also proposed by Camgöz et al. (2020a,b). For rendering sign language output, avatars are commonly used (Wolfe et al., 2022), as well as methods to generate videos of realistic signers (e.g. Saunders et al. 2022).

Parallel datasets In terms of datasets, past work in SLT can be characterized as focusing very much on a narrow linguistic domain, most of the work was done on one single data set called `RWTH-PHOENIX Weather 2014T` (Forster et al., 2014). PHOENIX has a size of 8k sentence pairs and contains only weather reports. The biggest parallel corpus for a European sign language to date, the Public DGS Corpus (Hanke et al., 2020), contains roughly 70k sentence pairs.

Thus, there is a clear shortage of usable parallel corpora, and existing ones are orders of magnitude smaller than what is considered an acceptable size for spoken language MT (as a rule of thumb, at least hundreds of thousands of sentence pairs). Nevertheless, there are plenty of spoken languages that also have little parallel data and MT methods have been developed specifically for low-resource MT (Sennrich and Zhang, 2019).

Evaluation For spoken language MT a variety of automatic metrics exist. These include more conventional, string-based metrics such as BLEU (Papineni et al., 2002) or chrF (Popović, 2015), as well as recent, learned metrics based on embeddings like COMET (Rei et al., 2020). In the context of SLT, no automatic metrics are validated empirically, but if the target language is spoken, many existing metrics are reasonable to use. However, if sign language is the target language, no automatic metric is known at the time of writing, and the only viable evaluation method is human evaluation. Apart from last year's shared task, a human evaluation of SLT systems has never been conducted on a large scale before, and there are open questions regarding the exact evaluation methodology and what the ideal profile (e.g. hearing status, language proficiency) for evaluators should be.

3 Tracks and submission procedure

We offered four translation directions ("tracks"): translation from DSGS to German and vice versa, French to LSF-CH, and Italian to LIS-CH.

For DSGS to German, submitted systems were ranked on a leaderboard. For all other directions, no automatic ranking was shown since automatic metrics of translation quality do not exist for sign languages as the target language.

We provided baseline systems for both translation scenarios (translating from or to a sign language). We were prepared to provide human evaluation for all submitted systems, regardless of the translation direction or language pair.

We deliberately did not limit the shared task to any particular kind of SL representation as input or output of an MT system. For DSGS-to-German translation, participants were free to use video frames, pose estimation, or something else. For German-to-DSGS participants were free to submit a video showing pose estimation output, an avatar, or a photo-realistic signer.

Participants had to submit their translation outputs on the OCELoT platform[5] which displayed an unofficial public leaderboard based on automatic metrics. Participants were allowed to make up to seven submissions and were asked to mark one of them as their primary submission.

Main outcome Four teams (including one from Northeastern University whose submission we consider a baseline) participated in our task. All of them submitted to the DSGS-to-German track, while there were no submissions for other translation directions.

4 Data

For this task we provided separate training, development and test data. While the training data was available from the beginning, the test data has been released in two stages, starting with a release of the test sources only.

Table 1 gives a high-level overview of our training, development and test data.

4.1 Licensing and attribution

Both datasets (SRF23 and Signsuisse) can be used for non-commercial research. Please note that distributing the datasets or making them accessible to third parties is not permitted, either in their original or edited form. In addition, this overview paper should be cited if the corpora are used.

4.2 Training Data

The training data comprises two corpora called Signsuisse (Jiang et al., 2023a) and SRF23 (Jiang et al., 2023b). Signsuisse is a multilingual dictionary containing lexical items in DSGS, LSF-CH and LIS-CH, represented as videos and glosses. Additionally, Signsuisse contains sentence-level parallel data as well, since there is one example sentence to show the use of the sign in context for each lexical item. SRF23 contains parallel data between DSGS and German, and its linguistic domain is general news. Both datasets are distributed through SwissUbase[6], where individual researchers had to agree with the usage terms and apply for access before downloading.

Training corpus 1: Signsuisse Lexicon We collected $18,221$ lexical items from the Signsuisse website, $17,221$ of which are released as training data and $1,000$ are reserved for testing and therefore not included in the training data release. The lexicon contains three languages: (i) DSGS (9044 items, 500 reserved), (ii) LSF-CH (6423 items, 250 reserved), and (iii) LIS-CH (2754 items, 250 reserved).

The lexical items are represented as videos and glosses, which enable sign-by-sign translation from spoken to signed languages. The videos were recorded with different framerates, either 24, 25, or 30 fps, and the video resolution is 640 x 480.

Training corpus 2: SRF23 These are daily national news and weather forecast episodes broadcast by the Swiss National TV (Schweizerisches Radio und Fernsehen, SRF)[7]. The episodes are narrated in Standard German of Switzerland (different from Standard German of Germany, and different from Swiss German dialects) and interpreted into Swiss German Sign Language (DSGS). The interpreters are hearing individuals, some of them children of Deaf adults (CODAs).

The subtitles are partly preproduced, and partly created live via respeaking to automatic speech recognition. While both the subtitles and the signing are based on the original speech (audio), due to the live subtitling and live interpreting scenario, a temporal offset between audio and subtitles as well as audio and signing is inevitable (Müller et al., 2022). It should also be pointed out that there are differences between interpreted and non-interpreted language (Dayter, 2019) due to source language interference and time constraints. SL during real-time interpretation tends to closely follow the grammatical structure of the spoken language (Leeson, 2005).

[5] https://ocelot-wmt23.mteval.org/

[6] https://www.swissubase.ch/en/catalogue/studies/20452/19280/overview

[7] https://www.srf.ch

		SRF23		Signsuisse		Total	
	direction	**episodes**	**segments**	**segments**	**lexical items**	**segments**	**lexical items**
training	DSGS↔DE	771	231834	9044	9044	240878	9044
	FR→LSF-CH	-	-	6423	6423	6423	6423
	IT→LIS-CH	-	-	2754	2754	2754	2754
development	DSGS↔DE	3	712	-	-	712	-
test	DSGS→DE	1	246	250	250	496	250
	DE→DSGS	1	258	250	250	508	250
	FR→LSF-CH	-	-	250	250	250	250
	IT→LIS-CH	-	-	250	250	250	250

Table 1: Overview of training, development and test data. SRF23 and Signsuisse are two different training corpora (§4.2). Segment count for the training corpora is after automatic sentence segmentation. The training data and development data for DSGS→DE and DE→DSGS are identical, while the test data is different. There was no designated development data for LSF-CH and LIS-CH.

Different from the first edition of the shared task (WMT-SLT22), the offset between the signing and the subtitles was not manually corrected for the training data of the current edition. On the other hand, the size of the training data is much larger than last year, presenting a different trade-off. See Table 2 for a comparison between this year's and last year's SRF resources. While last year our focus was providing training data of the highest quality, this year our focus was offering a large, noisy dataset that lends itself to data cleaning or filtering experiments such as automatic alignment.

Additional resources We encouraged participants to consider the MEDIAPI-SKEL corpus with parallel examples between French Sign Language and French (Bull et al., 2020a) as a further resource. Besides, we suggested that participants re-use the training corpora released for last year's shared task (Müller et al., 2022).

4.3 Development data

We did not provide any dedicated development data for this edition of the shared task. As is customary for WMT shared tasks, we encouraged participants to use last year's development and test data as development data for the current year.

4.4 Test data

We distribute separate test data for our four translation directions. See Table 1 for an overview.

DSGS→DE The test data consists of segments taken from undisclosed SRF23 and Signsuisse material (see §4.2 for a general description). The final test set is balanced, containing roughly 50% Signsuisse and 50% SRF23 examples. For the SRF23 part one episode was manually aligned using the iLex editor (Hanke and Storz, 2008), and the signer is a "known" person that appeared in the training set. We did not intend to test generalization to unknown signers during the shared task evaluation campaign. For the Signsuisse part we do not use the isolated lexical entries themselves for testing, but the example sentences associated with each lexical item.

DE→DSGS Same procedure as DSGS→DE, except that a different SRF23 episode and different sentences from Signsuisse are reserved for this translation direction.

FR→LSF-CH 250 undisclosed sentences from Signsuisse.

IT→LIS-CH 250 undisclosed sentences from Signsuisse.

5 Data preprocessing

For each data set described in §4 we provided videos and corresponding text in a spoken language. In addition, we included pose estimates (location of body keypoints in each frame) as a convenience.

5.1 Video processing (only SRF23)

Videos are re-encoded with lossless H264 and use an mp4 container. The framerate of videos is unchanged, meaning either 25, 30 or 50. We are not distributing the original videos but ones that are preprocessed in a particular way so that they only show the part of each frame where the signer is located (cropping) and the background is replaced with a monochrome color (signer masking), see Figure 1 for examples.

	SRF22	SRF23
Number of episodes	29	771
Time span of episodes	March 2020 to March 2021	July 2014 to May 2021
Total duration videos	16 hours	437 hours
Total number of subtitles (before/after sentence segmentation)	14265 / 7071	354901 / 231834
Number of signers	3	4
Subtitle segmentation	manual	automatic
Subtitle alignment	manual	audio

Table 2: Comparison between SRF training data of the 2022 and 2023 edition of the WMT-SLT shared task. Subtitle segmentation=ensuring that each subtitle unit is one entire sentence. Subtitle alignment=Subtitle times are either manually corrected to match the signing in the video (manual) or are matched with the audio track (audio).

Figure 1: Illustration of video preprocessing steps (cropping, instance segmentation and masking). From left to right: original frame, cropped frame, masked frame. Taken from Müller et al. (2022).

Cropping We manually annotate a rectangle (bounding box) around where the signer is located for each video. We then crop the video to only keep this region using the FFMPEG library.

Signer segmentation and masking To the cropped video we apply an instance segmentation model, Solo V2 (Wang et al., 2020), to separate the background from the signer. This produces a mask that can be superimposed on the cropped video to replace each background pixel in a frame with a grey color (`[127,127,127]` in RGB).

The video processing steps described above are only necessary for the SRF23 data, since Signsuisse footage is recorded against a neutral background and showing only one signer in the center of each frame.

5.2 Subtitle processing (only SRF23)

Since SRF23 subtitles are not manually aligned, automatic sentence segmentation[8] is used to redistribute text across subtitle segments, see Table 3 for examples. This process also adjusts timecodes in a heuristic manner if needed. For instance, if automatic sentence segmentation detects that a well-formed sentence stops in the middle of a subtitle, a new end time will be computed. The end time is proportional to the location of the last character of the sentence, relative to the entire length of the subtitle. See Example 2 in Table 3 for an illustration of this case.

5.3 Pose processing (both corpora)

"Poses" are an estimate of the location of body keypoints in video frames. The exact set of keypoints depends on the pose estimation system, well-known ones are OpenPose (Cao et al., 2019)[9] and MediaPipe Holistic (Lugaresi et al., 2019)[10]. Usually such a system provides 2D or 3D coordinates of keypoints in each frame, plus a confidence value for each keypoint.

The input for pose processing are cropped and masked videos (§5.1). See Figure 2 for examples of pose estimation on our data.

OpenPose We use the Openpose 137 model (which is the default) for the Signsuisse data and the Openpose 135 model for the SRF data. The two models are both widely used and the 137 model has two additional keypoints because it represents

[8] https://github.com/bricksdont/srt/tree/sentence_segmentation

[9] https://github.com/CMU-Perceptual-Computing-Lab/openpose

[10] https://ai.googleblog.com/2020/12/mediapipe-holistic-simultaneous-face.html

Example 1	
Original subtitle	After automatic segmentation
81 00:05:22,607 -> 00:05:24,687 Die Jury war beeindruckt 82 00:05:24,687 -> 00:05:28,127 und begeistert von dieser gehörlosen Frau.	48 00:05:22,607 -> 00:05:28,127 Die Jury war beeindruckt und begeistert von dieser gehörlosen Frau.
Example 2	
Original subtitle	After automatic segmentation
7 00:00:24,708 -> 00:00:27,268 Die Invalidenversicherung Region Bern startete 8 00:00:27,268 -> 00:00:29,860 dieses Pilotprojekt und will herausfinden, ob man es 9 00:00:29,860 -> 00:00:33,460 zukünftig umsetzen kann. Es geht um die Umsetzung	4 00:00:24,708 -> 00:00:31,720 Die Invalidenversicherung Region Bern startete dieses Pilotprojekt und will herausfinden, ob man es zukünftig umsetzen kann.

Table 3: Examples of automatic sentence segmentation for German subtitles. The subtitles are formatted as SRT, a common subtitle format. Taken from Müller et al. (2022).

Figure 2: Examples of the output of pose estimation systems overlaid over the original video frames. Left: OpenPose, right: MediaPipe Holistic. Taken from Müller et al. (2022).

the wrists twice. OpenPose often detects several people in our videos, even though there is only one single person present. We distribute the original predictions which contain all people that OpenPose detected.

MediaPipe Holistic As an alternative, we also estimate signers' poses with the MediaPipe Holistic system developed by Google. Unlike our OpenPose model, which only provides 2D joint locations, MediaPipe produces both 2D and 3D joint location coordinates. For the SRF data, values from Holistic are normalized between 0 and 1, instead of referring to actual video coordinates.

Unlike the first edition of the task, where the keypoints were stored in a JSON format, to deliver the pose data for more compact storage and faster I/O, in WMT-SLT 23 the binary .pose format of Moryossef and Müller (2021) was used.

6 Baselines and submitted systems

In this section we describe the submissions to our shared task. In case there are substantial differences between the primary and secondary submissions of a team we opted to describe the primary submission here. At the time of writing this overview paper three out of four teams have given us detailed information about their submissions. The submissions are summarized in Table 4.

Overall, the participating teams have diverse academic backgrounds, but their expertise is leaning towards NLP more than computer vision. All submitted systems are sequence-to-sequence models based on Transformers (Vaswani et al., 2017). Participants mostly chose to represent sign language data as video frames (using a visual feature extractor on the encoder side). Only the baseline system opted for Mediapipe pose features instead.

Two systems, by KNOWCOMP and TTIC, are unconstrained because their visual or spoken text components are pretrained on other datasets. Their approaches are best summarized as a combination of visual embeddings and pre-trained language models. TTIC used additional monolingual video data from OpenASL for pretraining, and no submission used monolingual text in a spoken language.

Two teams have published their code, with another team planning to do so in the future.

6.1 Baseline by Northeastern University (DSGS→DE)

Based on the models of the previous challenge, we pre-train the baseline signed-to-spoken system using a Transformer architecture. We use the fairseq seq2seq translation library (Ott et al., 2019), and the open-source implementation of the architecture by Tarrés et al. (2023). We first train a Sentencepiece tokenization model on the German text of the example sentences of the Signsuisse dataset. Then, we train the model on the Mediapipe Holistic poses on the Signsuisse example sentences. We, then, validate and test the model on the extracted Mediapipe Holistic poses of both the Signsuisse and SRF DSGS-to-German datasets. The final output is detokenized to result in spoken German text.

6.2 Baseline by UZH (DE→DSGS, FR→LSF-CH, IT→LIS-CH)

As a naive solution, we choose a sign-by-sign translation baseline (Moryossef et al., 2023). The system gets German text as input, performs text-to-gloss translation, then for each gloss looks up a sign in the Signsuisse lexicon. The estimated poses from each sign are then concatenated and smoothed out, to create a single pose video with the translation into a sign language.

Since there were no submissions by participants to these tracks, this baseline was not used for any subsequent evaluation.

6.3 Submission by KNOWCOMP (Xu et al., 2023)

The team proposed a framework which combines a pre-trained visual model to extract visual embeddings with a GPT2-based language model to translate into text.

The framework first utilises an I3D model (Varol et al., 2022) pre-trained on the BSL-1K corpus (Albanie et al., 2020) to extract 1024-dimensional tensors for a 64-frame video input. The video extractor, i.e. the I3D model, generates a 1024-dimensional tensors as the visual representation of the input video (64 frames). For decoding, a German-GPT2 (Radford et al., 2019) large language model (LLM) is used to generate the final translations. To establish an alignment between the visual and the textual embeddings from the two models, the team trains an embedding alignment block to project the obtained visual embeddings into textual embeddings.

	BASELINE	KNOWCOMP	TTIC	CASIA
Constrained	✔	-	-	?
Multilingual	-	-	-	?
Document-level	-	-	-	?
Model ensemble	-	-	-	?
Pretrained components	-	✔	✔	?
Monolingual data	-	✔	✔	?
Synthetic data	-	-	-	?
Signed language representation	Mediapipe	I3D features	Video frames	?
Spoken language representation	SP	BPE	SP	?
Open-source code	✔	(✔)	✔	?

Table 4: Overview of characteristics of submitted systems. CASIA did not disclose any information. In the code row, checkmarks are clickable links. BPE=Byte Pair Encoding, SP=Sentencepiece, (✔)=authors plan to publish the code.

This is implemented by stacking 6 Transformer encoder layers together. Two fully connected neural networks are placed before and after the alignment block to extend the visual embeddings into a sequential format and to densify the aligned embeddings into prefix embeddings for German-GPT2, respectively.

Before training their model KnowComp first employs a data preprocessing step where the raw data is divided into smaller video segments which are then matched with the corresponding ground truth German translations. To ensure that the input observes the visual model requirements, i.e. input of 64 frames, they downsample the video segments taking the first of each three frames. In cases where the video segment is smaller than 64 frames, pure black frames are appended. Next, the video frames are resized to 224 x 224.

At training time, to enhance training efficiency, the parameters of the visual and the translation models are first frozen; later, at a certain iteration, the parameters of GPT2 are unfrozen. This strategy ensures that the randomly initialized Transformer encoder does not compromise the LLM. The hyperparameters they used are: batch size of 4, learning optimizer Adam (Kingma and Ba, 2015) with a learning rate of $5e-6$, and unfreezing the training parameters at iteration 66000. The input and output lengths of GPT2 were set to 20. The number of heads in the multi-head attention was set to 8; the prefix length for GPT2 to 4. Before the visual embeddings were fed to the alignment block, the sequence length was adjusted to 2×4, where 4 is the GPT2's prefix number. They ran their experiments on an NVIDIA GeForce GTX 1080 Ti with 11G VRAM.

6.4 Submission by TTIC (Sandoval-Castaneda et al., 2023)

The system by the TTIC team uses as visual backbone the VideoSwin Transformer (Liu et al., 2022) and the T5 model by Raffel et al. (2020) for translation into text. The VideoSwin model was pretrained on the visual (video) side of OpenASL (Shi et al., 2022, thus excluding the English translations) using the codebook from a discrete variational auto-encoder (dVAE, Ramesh et al., 2021) to produce the labels in the self-supervision objective. Next, the model was fine-tuned for the task of isolated sign language recognition on the gloss-based version (Dafnis et al., 2022) of the WLASL2000 dataset (Li et al., 2020).

The input data was segmented into non-overlapping, padded chunks of 16 frames in order to meet the input requirements of VideoSwin. The outputs were concatenated together.

Following the findings of Uthus et al. (2023) that English pre-trained T5 and fine-tuned for ASL to English translation produces state-of-the-art results, the TTIC team used a T5 model pre-trained on the German Colossal Cleaned Common Crawl (GC4) corpus.[11] They used pre-trained checkpoints from HuggingFace (Wolf et al., 2019). To tokenize the target side, SentencePiece (Kudo and Richardson, 2018) trained on the same data was used to produce a vocabulary of 32,128 tokens.

Their system employs a convolutional layer that is trained to project the sequence of visual features into a single vector per time step. The T5 embeddings layer is replaced by this convolutional layer. The cross-entropy loss was used for the BEVT pre-

[11] https://german-nlp-group.github.io/projects/gc4-corpus.html

training, the ISLR fine-tuning, the text-to-text pre-training as well as for the translation. At inference time, the diverse beam search algorithm (Vijayakumar et al., 2016) with 5 beams, 5 beam groups and a diversity penalty of 1 was used. In contrast to KNOWCOMP, the TTIC team used 8 GPUs to train their system.

6.5 Submission by CASIA

Finally, we received several submissions from the National Laboratory of Pattern Recognition at the Institute of Automation, Chinese Academy of Sciences (submission ID: CASIA). No system paper was submitted and the authors did not provide further information.

7 Evaluation Protocols

We performed both a human (§7.1) and an automatic (§7.2) evaluation of translation quality. Our final system ranking is based on the human evaluation only.

7.1 Human evaluation

Our human evaluation follows the setting we established last year for SLT human evaluation with custom guidelines (Müller et al., 2022), which was originally adapted from the evaluation protocol used at the recent WMT conferences (Kocmi et al., 2022).

Scoring method We employed the source-based direct assessment (DA; Graham et al., 2013; Cettolo et al., 2017) methodology with document context, extended with Scalar Quality Metric (SQM; Freitag et al., 2021). Assessments were performed on a continuous scale between 0 and 100 as in traditional DA but with 0-6 markings on the analogue slider and custom annotator guidelines specifically designed for our task.

As a result of the human evaluation, the systems are ranked from best to worst, after averaging the segment-level DA scores given by the human annotators. In contrast to previous evaluation campaigns (Akhbardeh et al., 2021) which calculate the rankings based on standardized scores (z-scores), we decided to not do so, because the large number of zero-scored items led to a rather skewed standardization scale which affected the calculation of the clusters. We did not make any distinction between segment-level and document-level scores, simply including the latter as additional data for computing the average scores.

After ranking the systems based on their average scores, they are grouped into significance clusters, following the Wilcoxon rank-sum test. Rank ranges give an indication of the translation quality of a system within a cluster and are based on the same head-to-head statistical significance tests.

Inter- and intra-annotator agreement was measured with Fleiss κ (Fleiss, 1971). This should be considered an approximation, noting the concerns of Ma et al. (2017) that kappa coefficients are not suitable for continuous scales. In order to calculate the coefficient, the values have been discretized in seven bins in the scale 0-6, since those were the scores marked on the continuous evaluation bar that was given to the annotators.

Settings of evaluation campaign We used the Appraise evaluation framework[12] (Federmann, 2018) for collecting segment-level judgments. As there were submissions in the DSGS-to-German direction only (§6), we only set up a sign-to-text human evaluation campaign. Annotators were presented with video fragments as source context and translation outputs of a random document fragment from an MT system. The reference translation and the official baseline were included as additional system outputs. Document fragments were created from (up to) twelve consecutive segments. The SRF23 part of the test set was evaluated within the document context. Because the Signsuisse part is a collection of utterances without document boundaries, we presented up to twelve random segments at once but emphasized in the guidelines that those are unrelated and should be assessed independently.

A screenshot of an example annotation in Appraise is presented in Figure 3. The full instructions to evaluators in English and German are listed in Appendix B.

Data and scripts used for generating tasks and computing the final system rankings are publicly available in a Github repository.[13]

We hired three evaluators who are native German speakers and trained DSGS interpreters. All of them had prior experience with evaluation of MT output. Each evaluator was assigned an identical set of annotation tasks comprising the entire test set and all participating systems, including the baseline system and the reference translation. As last year, we did not include any quality control items in the annotation tasks as we had multiple independent

[12] https://github.com/AppraiseDev/Appraise
[13] https://github.com/WMT-SLT/wmt-slt23

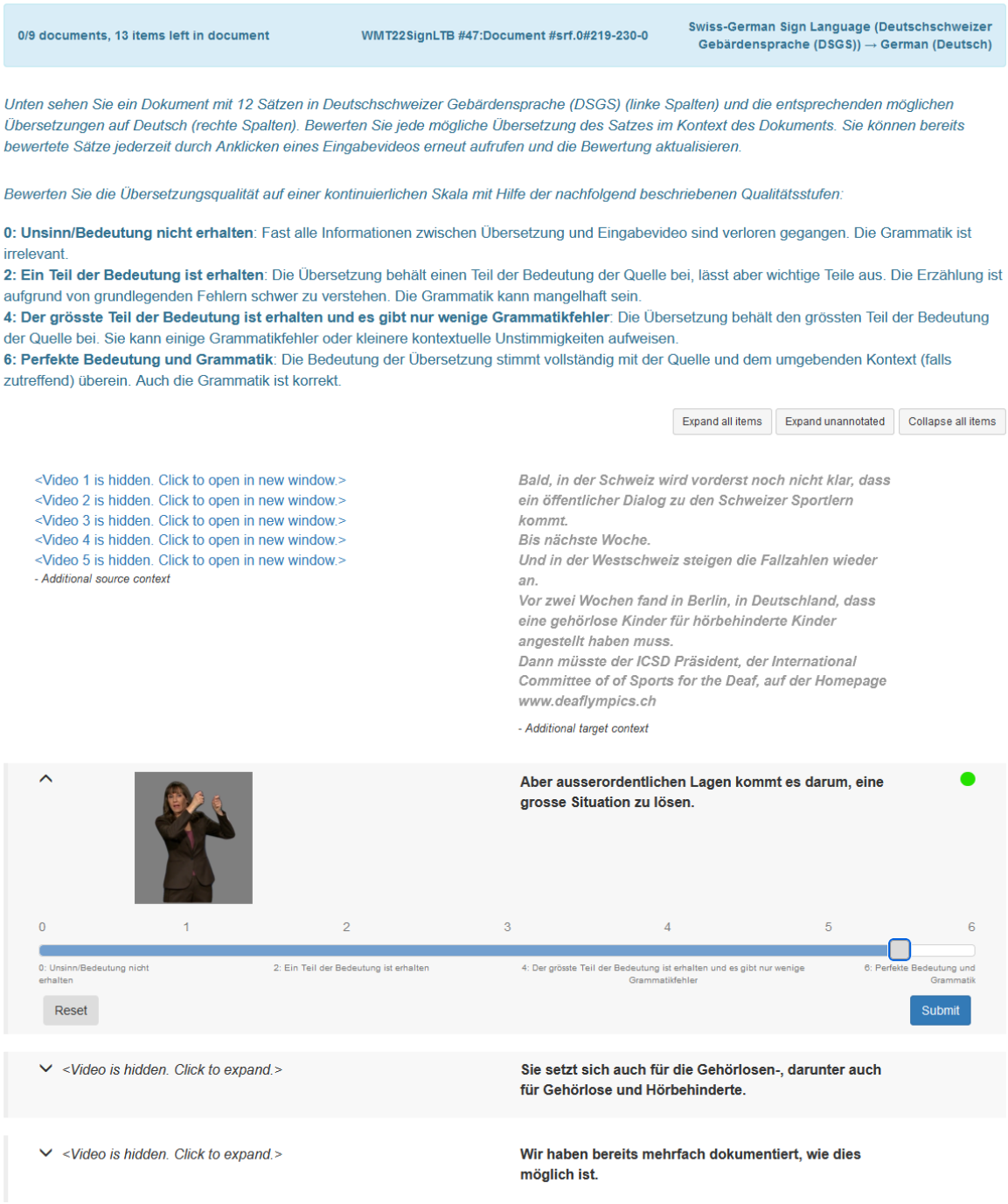


Figure 3: A screenshot of an example sign-to-text annotation task in Appraise featuring document-level source-based direct assessment (DA) with scalar quality metrics (SQM) and custom annotator guidelines in German. Taken from Müller et al. (2022).

annotations of the entire test set and because of the very low quality of translations, which would make them indistinguishable from segments with randomly replaced words or phrases used as quality control items.

Feedback from evaluators After completing the evaluation all three evaluators filled out the feedback form we used last year regarding the evaluation procedure and the Appraise platform, where they gave us additional informal feedback.

7.2 Automatic evaluation

As in the previous edition, to complement our human evaluation (which provides the main ranking) we also provide an automatic evaluation. We evaluate the submissions from DSGS into German using three automatic metrics: BLEU (Papineni et al., 2002), chrF (Popović, 2015) and BLEURT (Sellam et al., 2020). We note that learned, semantic metrics correlate better with human judgement (Kocmi et al., 2021), but if they consider the source text as an input (e.g. COMET; Rei et al., 2020), they cannot be used in our context because our source is video and not text. There is no known learned metric which supports sign language videos. We use sacreBLEU (Post, 2018) for BLEU[14] and chrF[15] and the Python library for BLEURT.[16] In all cases, we estimate 95% confidence intervals via bootstrap resampling (Koehn, 2004) with 1000 samples.

8 Results

8.1 Human evaluation

Assessment scores All three evaluators completed all tasks, which gave us three independent judgements for each segment from the official test set. In total, for the output of five systems, we collected 7,800 segment-level and 792 document-level assessment scores, which averages to 1,718 scores per system.

System ranking The official system ranking is presented in Table 5. The significance clusters are indicated with horizontal lines. According to our human evaluation (Table 5), the submission by TTIC has achieved an average score of 0.7 on the scale of 0 to 100, compared to a score of 83.8 for human translations. The score of TTIC is significantly better than the other systems in the table. All other systems ended up in the same cluster with overall lower translation quality.

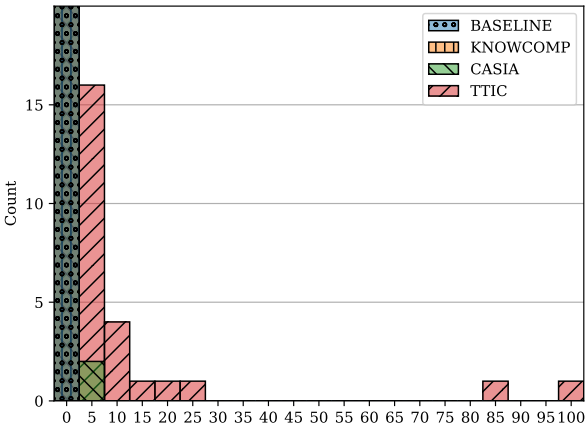


Figure 4: Histogram with the distribution of the system outputs at the DA score scale (x axis) with overlapping semi-transparent bars, discretized into 20 bins. For every segment we include only the average of all ratings. Bin 0, where most ratings belong (up to 496), is cropped to 20 to make the histogram visible.

Distribution of scores In order to make the distribution of DA scores more interpretable, it is visualized in Figure 4. TTIC had one segment with a score of 99 out of 100, one with 83, one for each of the scores 22, 18 and 15, then 4 segments with a score of about 10, and 16 segments with a score of about 5. CASIA had two segments with a score of about 5. The rest of the segments, including all the outputs from the KNOWCOMP and BASELINE systems, have been given a score very close to 0.

Some example outputs of the highest-scoring translations are listed in Table 6. One can see that TTIC came close to correctly translating the general introductory greetings of the news, but for the rest of the MT ouputs, rated less than 20 out of 100, only a few words match the reference.

Annotator agreement In Table 7 we are reporting intra-annotator agreement for every annotator, measured with Fleiss κ (Fleiss, 1971) over 134 segments which were evaluated twice. (Landis and Koch, 1977; Agresti, 1996). The inter-annotator agreement is $\kappa = 0.80 \pm 0.01$. One can observe that the intra-annotator agreement and all 3 intra-annotator agreements are substantial ($0.61 < \kappa \leq 0.80$) based on Landis and Koch, 1977).

[14] `BLEU|nrefs:1|bs:1000|seed:12345|case:mixed|eff:no|tok:13a|smooth:exp|version: 2.2.0`

[15] `chrF2|nrefs:1|bs:1000|seed:12345|case:mixed|eff:yes|nc:6|nw:0|space:no|version: 2.2.0`

[16] BLEURT v0.0.2 using checkpoint BLEURT-20.

both domains		
Rank	Ave.	System
1	83.829	HUMAN
2	0.669	TTIC
3-5	0.024	CASIA
3-5	0.008	BASELINE
3-5	0.005	KNOWCOMP

SRF		
Rank	Ave.	System
1	68.809	HUMAN
2	1.192	TTIC
3-4	0.046	CASIA
3-5	0.009	BASELINE
4-5	0.002	KNOWCOMP

Signsuisse		
Rank	Ave.	System
1	98.630	HUMAN
2	0.154	TTIC
3-5	0.008	BASELINE
3-5	0.007	KNOWCOMP
3-5	0.003	CASIA

Table 5: Official results of the WMT23 Sign Language Translation task for translation from Swiss German Sign Language to German. Systems are ordered by averaged (non-standardized) human score in the percentage scale. Lines indicate clusters according to a Wilcoxon rank-sum test $p < 0.05$.

score	**system**	**testset**	**doc**	**seg**		**text**
99.3	TTIC	SRF	0	0	hyp:	Guten Abend, meine Damen und Herren, willkommen zur "Tagesschau".
					ref:	Guten Abend, meine Damen und Herren, willkommen zur "Tagesschau".
83.3	TTIC	SRF	0	1	hyp:	Heute mit diesen Themen:
					ref:	Das macht heute Montag Schlagzeilen:
18.7	TTIC	SRF	23	9	hyp:	Der US-Präsident ist heute zu Gast bei "10vor10".
					ref:	Wesentliches gibt es auch heute bei "10vor10".
16.3	TTIC	SRF	18	0	hyp:	Und auch für EU-Bürger, die in die Schweiz einreisen wollen, soll es verschärfte Einreiseregeln geben.
					ref:	Auch die EU will nun ihre Bürger vom Kreuzfahrtschiff zurückholen, denn man misstraut Japans Krisenmanagement.
12.0	TTIC	SRF	14	2	hyp:	Die Leute müssen sich Gedanken machen, wie sie die Zukunft meistern können.
					ref:	Das muss sich ändern, sind sich die EU-Aussenminister einig.
11.0	TTIC	SS	18	5	hyp:	Der Film kann auf YouTube angeschaut werden.
					ref:	Dieser Film ist spannend und interessant.
8.3	TTIC	SRF	15	4	hyp:	Tausende Menschen sind seither ohne Hilfe von aussen ausgewandert.
					ref:	Über 70'000 Menschen haben sich bis heute mit dem neuen Coronavirus infiziert.
5.0	CASIA	SRF	1	1	hyp:	Die Temperaturen steigen in der Schweiz.
					ref:	Und morgen gibt es sonnige Phasen bei Temperaturen um 9 °C.

Table 6: Examples of some of the highest-scoring translations in the test set. hyp=MT outputs, ref=human translation

annotator	kappa
A	0.80±0.05
B	0.80±0.06
C	0.79±0.06

Table 7: Intra-annotator agreement based on the Fleiss κ coefficient for reliability of agreement (with scores discretized in the scale 0-6).

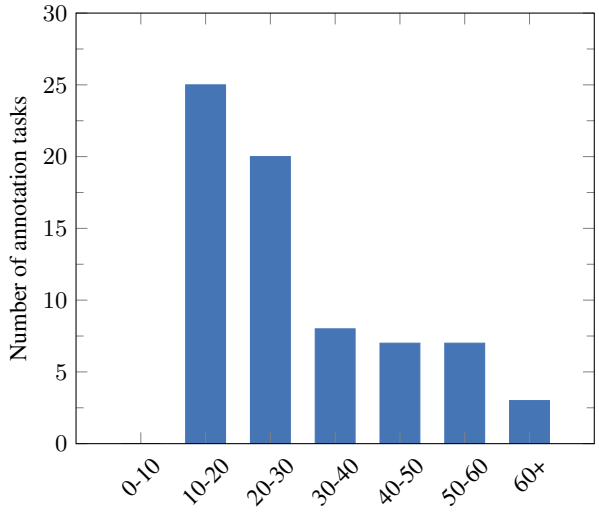


Figure 5: Number of task completion times (a task consists of 100 segments) grouped into 10-minute buckets, after removing top and bottom 5-percentiles.

Evaluation speed A single task requiring providing 100 segment-level and about 12 document-level scores took on average 29 minutes to complete, after excluding 5% of slowest and fastest task annotations. The majority of tasks were finished in between 10 and 30 minutes as shown in Figure 5. This is substantially faster than last year, which averaged around 45 minutes per task.

Feedback from evaluators After completing the evaluation all three evaluators filled in a form meant for feedback regarding the evaluation procedure and the Appraise platform. All evaluators gave us additional informal feedback.

In general, evaluators reported that their experience with Appraise was positive (two of them had used Appraise before), and that our instructions were clear. All of them would be willing to do similar work in the future. They found source videos understandable and the documents or segments given were neither too long nor too short. The general method of assessing translations (DA with SQM) was not found difficult nor stressful, but on the contrary annotators thought it was efficient, simple, fast and practical.

Concerning Appraise development, nobody experienced technical problems, which is an improvement over last year, when two people experienced major technical issues. Evaluators suggested that the user interface could be improved in some places. For instance, automatically playing videos could make evaluations more efficient, the videos should be bigger by default, there should be more keyboard shortcuts and there should be a quick way to give a low score to an entire document.

As explained in more detail below (§9.3), and similar to last year, evaluators told us that some videos do not have ideal cuts, in the sense that the beginning or end are slightly cut off. This is perhaps inevitable in continuous signing, or a problem in our manual alignment process.

Full responses to the feedback form submitted by evaluators are listed in Appendix C.

8.2 Automatic evaluation

Table 8 summarises the results of the automatic evaluation. In general, the translation of the Signsuisse subset (SS) and the SRF23 subset seem to have a similar complexity, especially according to chrF and BLEURT evaluation scores. BLEU, on the other hand, shows higher translation quality for SRF in selected systems by CASIA and TTIC. Both teams are able to significantly outperform the baseline system according to the three evaluation metrics. TTIC achieves the best scores with their primary submission TTIC.423. Although chrF points out another of their submissions as the best system, the difference with respect to the primary submission is not statistically significant.

9 Discussion

9.1 General translation quality

Overall, all systems perform poorly in our shared task, as there is an extreme difference in average score between all systems and the human reference translation. The systems exhibit well-known problems of natural language generation such as overfitting to few high-probability hypotheses and hallucination (Lee et al., 2018; Raunak et al., 2021).

The best submitted system in the best case achieves an average score of about 1 out of 100 (where the human translation achieved 69 out of 100), which indicates that current automatic translations are not usable in practice, unlike spoken language MT where in specific scenarios experiments have shown systems to be on par with human

Submission	BLEU			chrF			BLEURT		
	all	SS	SRF23	all	SS	SRF23	all	SS	SRF23
BASELINE	0.09±0.03	0.15±0.06	0.10±0.05	12.4±0.4	12.2±0.5	12.5±0.5	0.072±0.003	0.083±0.005	0.060±0.005
CASIA.426	0.38±0.20	0.16±0.04	0.52±0.28	14.6±0.4	14.2±0.5	14.8±0.7	0.148±0.006	0.143±0.008	0.152±0.007
CASIA.427	0.39±0.20	0.13±0.05	0.52±0.28	14.2±0.5	13.4±0.5	14.8±0.7	0.162±0.006	0.171±0.009	0.152±0.007
CASIA.428	0.16±0.07	0.16±0.04	0.20±0.10	13.5±0.4	14.2±0.5	13.0±0.5	0.156±0.005	0.143±0.008	0.168±0.007
CASIA.429	0.38±0.20	0.15±0.06	0.52±0.28	14.3±0.4	13.5±0.5	14.8±0.7	0.175±0.006	0.197±0.008	0.152±0.007
CASIA.430	0.33±0.16	0.15±0.10	0.52±0.28	14.7±0.4	14.6±0.5	14.8±0.7	0.166±0.006	0.179±0.008	0.152±0.007
CASIA.431	0.13±0.06	0.15±0.10	0.14±0.03	14.5±0.4	14.6±0.5	14.4±0.6	0.169±0.006	0.179±0.008	0.159±0.008
CASIA.432	0.37±0.19	0.11±0.05	0.52±0.28	14.4±0.4	13.7±0.5	14.8±0.7	0.172±0.006	0.190±0.008	0.152±0.007
KNOWCOMP.418	0.06±0.03	0.07±0.03	0.09±0.04	6.2±0.3	6.9±0.5	5.7±0.5	0.077±0.005	0.080±0.007	0.073±0.007
KNOWCOMP.419	0.07±0.05	0.06±0.02	0.11±0.09	7.6±0.3	8.2±0.4	7.2±0.4	0.083±0.005	0.084±0.007	0.081±0.007
TTIC.417	0.56±0.46	0.30±0.14	0.29±0.13	15.9±0.5	16.6±0.8	15.3±0.6	0.222±0.010	0.231±0.011	0.210±0.015
TTIC.420	0.78±0.83	0.21±0.04	0.17±0.02	16.0±0.5	16.2±0.6	15.5±0.6	0.224±0.010	0.228±0.011	0.216±0.015
TTIC.421	0.21±0.09	0.13±0.06	0.29±0.13	13.2±0.4	13.3±0.5	13.2±0.6	0.087±0.006	0.078±0.006	0.095±0.010
TTIC.422	0.77±0.74	0.22±0.13	0.29±0.12	17.3±0.5	16.7±0.6	17.4±0.6	0.239±0.010	0.230±0.011	0.245±0.015
TTIC.423	1.03±0.87	0.21±0.03	0.69±0.46	17.0±0.6	16.2±0.7	17.2±0.7	0.243±0.010	0.236±0.011	0.246±0.013
TTIC.424	0.79±0.74	0.24±0.12	0.33±0.14	17.2±0.5	16.6±0.7	17.5±0.7	0.236±0.009	0.228±0.011	0.241±0.015
TTIC.425	0.74±0.79	0.14±0.06	0.23±0.10	16.3±0.6	16.0±0.7	16.3±0.7	0.205±0.009	0.194±0.010	0.214±0.014

Table 8: Automatic evaluation of all the submission for the full WMT-SLT test set (all), the Signsuisse subset (SS) and the SRF23 subset. Mean and 95% confidence intervals obtained via bootstrap resampling are shown. Primary submissions manually evaluated are boldfaced.

translation (Hassan et al., 2018; Popel et al., 2020). This assessment of general translation quality is unchanged from last year, see Müller et al. (2022) for potential reasons that still apply to the current shared task.

9.2 No submissions for spoken-to-signed translation directions

No teams participated in a track where a sign language is the target language (§3). We believe this could be due to the fact that generating sign language may appear considerably harder to participants. The problem of signed-to-spoken translation fits well into existing translation paradigms and toolkits, because using arbitrary features on the source side is easier than generating arbitrary numerical data (such as a video). Decoding text on the target side is considerably easier and more well-defined in NLP than decoding a video or similar data structure.

We thought that providing a baseline system for spoken-to-signed translation (§6.2) may help lower the barriers to entry but clearly, more measures are needed. A different hypothesis is that our shared task in its current form does not appeal to scientists working in the field of sign language generation or avatar technology. They may have felt alienated by aspects of the shared task which are familiar to MT researchers, but would need more explanation or introduction for people from neighboring fields.

9.3 Low scores of human translations

When looking at the domain-specific results (Table 5b and c), we observe that the human translation in SRF was ranked considerably lower than Signsuisse (69% against 98%). This difference warrants further investigation, as does the fact that a percentage of 69% is by itself rather low. We explain potential reasons for this below, attributing the difference to the way the corpora were generated.

Interpretation vs. translation SRF is partially generated as live interpretation of the spoken TV shows (spoken-to-sign), where interpreters are under time pressure. Due to specific efficiency strategies they occasionally omit content to keep up with the spoken audio. Therefore, since here we are evaluating the performance of the systems in the opposite direction (sign-to-spoken) it may as well very often be that the content of the interpretation does not match the one of the written or spoken sentence. However, as explained in Section 4, the Signsuisse part of the testset derives from a lexicon, containing sentences recorded as examples of particular lexicon entries. Since these have been generated for the purpose of being included in the lexicon, the accuracy of the translation is expected to be much higher than the one achieved within live interpretation.

Video editing issues The measured bad human performance on SRF may also be explained by the fact that the video cuts are sometimes not ideal,

i.e. the beginning or end of an SL utterance is cut off, as noted by our evaluators. This may have occurred because segmenting continuous signing is difficult and there is no ideal way to separate seamless transitions.

In the future these problems could perhaps be mitigated by including more frames from the left and right border of a video clip, or simply discarding sentences with unclear boundaries.

Role of discourse context A third reason may be that SLs are probably more dependent on context than spoken languages, e.g. because of index signs. This means that evaluating an isolated SL utterance (the equivalent of one sentence in a spoken language) may lead to low scores. This is a phenomenon that would more likely occur in a news report of SRF, as compared to the isolated example sentences of Signsuisse.

Contrary to what was observed for the evaluation of the human translation, the two submitted MT systems TTIC and CASIA perform significantly better on SRF than on Signsuisse. Here we may provide the assumption, that since the amount of training sentences from SRF is bigger than the ones from Signsuisse, the systems are optimized better for that domain. Additionally, it has been noted that in interpretation settings similar to the ones of SRF, the linguistic characteristics of the signing may be more closely related to German than in an offline translation setting, such as the one in Signsuisse.

9.4 Quality of training data and unexplored potential

Compared to last year we offered considerably more training data (hundreds of hours worth of video compared to dozens last year; §4.2). However, while last year all training data was manually corrected, this year we offered the data as-is. The SRF23 training data is best understood as a comparable corpus, or web-crawled parallel corpus including various types of noise (Khayrallah and Koehn, 2018). For instance, the time stamps of the German subtitles are more aligned with the audio signal present in the broadcast and do not account for the delay of live-interpreted signing. Any naive extraction of parallel examples from SRF23 without any alignment tools or shifting subtitle times will result in noisy training data.

As far as we know no participant investigated ways to improve the alignments automatically, which is perhaps because we did not explain this well in our online documentation. One reason for this may be that we did not make it clear enough to participants that one of our training corpora is effectively un-aligned. But essentially, it means there is unexplored potential in improving or filtering the training data instead of training on the raw corpora.

9.5 Limitations of shared task setup

The limitations we identified in last year's findings paper still apply. Briefly, the limitations concern the lack of generalization across signers, the favourable recording conditions of our sign language data and interpretation vs. translation setups. See Müller et al. (2022) for a more comprehensive description.

10 Conclusion and future directions

In this paper we present the second WMT Shared Task on Sign Language Translation (WMT-SLT23). We consider automatic sign language translation, and sign language processing in general, to be of wide public interest and to have a high potential impact in a societal and academic sense (§2).

Compared to last year we ran our shared task for three language pairs instead of one, we distributed considerably more training data (albeit with a higher amount of noise) and we put more emphasis on scenarios where sign languages are the target language.

Four teams participated in the second edition of the shared task. Overall, we observed low system performance with an average human evaluation score of about 1 out of 100 (for the best-performing system), which is not usable in practice. The main reasons for this outcome are a lack of usable training data, a modality gap (considering that most existing work in MT is based on text) and a lack of basic NLP tools specifically for sign languages.

Future of the shared task After two successful iterations the shared task is now well established, in the sense that suitable protocols are in place for human and automatic evaluation, reasonable baseline systems exist, as well as several training corpora and official WMT test sets.

So far our shared tasks have certainly helped to paint a more realistic picture of the translation quality of state-of-the-art systems, but they have not led to any major technical innovation. This may be because technologies more fundamental than machine translation do not exist for sign languages, or are not reliable enough. For this reason we will

consider running shared tasks on more fundamental problems in SLP such as alignment, segmentation, or automatic filtering of parallel corpora.

In the future we could also try to shift the focus away from interpreted news broadcast material as the basis for training and test data. A major challenge to overcome is that interpreted material is available in larger amounts, while signing produced by conventional, off-line translation or produced by native signers is harder to come by. Nevertheless, using non-interpreted material largely avoids alignment shifts in the training data and leads to higher scores for the human translations of the test data, among other advantages.

11 Ethical statement

Within this shared task, two main ethical considerations emerge: the potential impact of SL technology on target users and privacy considerations.

Research in sign language processing, if not executed carefully, may inadvertently cause harm to end users, especially members of deaf communities. Hearing scientists should refrain from prescribing what sort of language technology should be accepted by deaf or hard-of-hearing individuals and should avoid claiming that their approach "solves" any particular problem. Ideally, research of this nature should include deaf and hard-of-hearing people, not only at evaluation time but in the entire development cycle (Fox et al., 2023).

Secondly, there is a concern for the privacy of individuals depicted in SLP datasets. For the specific use case of sign language data, proper anonymisation is impossible, since identifying details such as facial expressions are crucial for sign language communication. We have obtained written permission of all individuals shown in our datasets. Storing and processing pose estimation features instead of raw videos may be an alternative that provides anonymity (and has other generalization effects such as ignoring differences in race, gender, clothing, background, etc.). However, in our shared task and related literature, (Moryossef et al., 2021; Tarrés et al., 2023) video features outperform pose features.

Acknowledgements

The organizing committee acknowledges funding from the following projects: the EU Horizon 2020 projects EASIER (grant agreement number 101016982) and SignON (101017255), the Swiss Innovation Agency (Innosuisse) flagship IICT (PFFS-21-47) and the German Ministry of Education and Research (BMBF) through the project SocialWear (01IW20002).

Finally, we would like to extend heartfelt thanks to the DSGS interpreters who performed our human evaluation: Heidi Stocker, Janine Criblez and Tanja Joseph.

References

Alan Agresti. 1996. *An introduction to categorical data analysis*, volume 135. Wiley New York.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 Conference on Machine Translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. 2020. Bsl-1k: Scaling up co-articulated sign language recognition using mouthing cues. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 35–53. Springer.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-Scale Acquisition of Parallel Corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Elise Bertin-Lemée, Annelies Braffort, Camille Challant, Claire Danet, and Michael Filhol. 2023. Example-based machine translation from textto a hierarchical representation of sign language. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 21–30, Tampere, Finland. European Association for Machine Translation.

Claudia Bianchini and Fabrizio Borgia. 2012. Writing sign languages: analysis of the evolution of the signwriting system from 1995 to 2010, and proposals for future developments. In *Proceedings of the Intl Jubilee Congress of the Technical University of Varna*, pages 118–123.

Danielle Bragg, Oscar Koller, Mary Bellard, Larwan Berke, Patrick Boudreault, Annelies Braffort, Naomi Caselli, Matt Huenerfauth, Hernisa Kacorri, Tessa Verhoef, Christian Vogler, and Meredith Ringel Morris. 2019. Sign Language Recognition, Generation, and Translation: An Interdisciplinary Perspective. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, pages 16–31, New York, NY, USA. Association for Computing Machinery.

Hannah Bull, Annelies Braffort, and Michèle Gouiffès. 2020a. Mediapi-skel-a 2d-skeleton video database of french sign language with aligned french subtitles. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 6063–6068.

Hannah Bull, Michèle Gouiffès, and Annelies Braffort. 2020b. Automatic segmentation of sign language into subtitle-units. In *European Conference on Computer Vision*, pages 186–198. Springer.

Necati Cihan Camgöz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural Sign Language Translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7784–7793.

Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020a. Multi-channel transformers for multi-articulatory sign language translation. In *European Conference on Computer Vision*, pages 301–319.

Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020b. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10023–10033.

Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Niehues Jan, Stüker Sebastian, Sudoh Katsuitho, Yoshino Koichiro, and Federmann Christian. 2017. Overview of the IWSLT 2017 evaluation campaign. In *International Workshop on Spoken Language Translation*, pages 2–14.

Onno Crasborn. 2006. Nonmanual structures in sign language. *Encyclopedia of Language and Linguistics*, 8:668–672.

Konstantinos M Dafnis, Evgenia Chroni, Carol Neidle, and Dimitri Metaxas. 2022. Bidirectional skeleton-based isolated sign recognition using graph convolutional networks. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7328–7338.

Daria Dayter. 2019. Collocations in non-interpreted and simultaneously interpreted english: a corpus study. In *New empirical perspectives on translation and interpreting*, pages 67–91. Routledge.

Mathieu De Coster, Dimitar Shterionov, Mieke Van Herreweghe, and Joni Dambre. 2022. Machine translation from signed to spoken languages: State of the art and challenges. *arXiv preprint arXiv:2202.03086*.

Mirella De Sisto, Dimitar Shterionov, Irene Murtagh, Myriam Vermeerbergen, and Lorraine Leeson. 2021. Defining Meaningful Units. Challenges in Sign Segmentation and Segment-Meaning Mapping (short paper). In *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 98–103, Virtual. Association for Machine Translation in the Americas.

Mirella De Sisto, Vincent Vandeghinste, Santiago Egea Gómez, Mathieu De Coster, Dimitar Shterionov, and Horacio Saggion. 2022. Challenges with sign language datasets for sign language recognition and translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2478–2487, Marseille, France. European Language Resources Association.

Elisabeth Engberg-Pedersen. 1993. *Space in Danish Sign Language: The Semantics and Morphosyntax of the Use of Space in a Visual Language*. SIGNUM-Press.

Christian Federmann. 2018. Appraise Evaluation Framework for Machine Translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.

Michael Filhol. 2020. Elicitation and Corpus of Spontaneous Sign Language Discourse Representation Diagrams. In *Proceedings of the 9th Workshop on the Representation and Processing of Sign Languages at Language Resources and Evaluation Conference*, pages 53–60. European Language Resources Association (ELRA).

Joseph L. Fleiss. 1971. Measuring Nominal Scale Agreement Among Many Raters. *Psychological bulletin*, 76(5):378.

Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. 2014. Extensions of the Sign Language Recognition and Translation Corpus RWTH-PHOENIX-Weather. In *Proceedings of the Ninth International Conference on Language*

Resources and Evaluation (LREC'14), pages 1911–1916, Reykjavik, Iceland. European Language Resources Association (ELRA).

Neil Fox, Bencie Woll, and Kearsy Cormier. 2023. Best practices for sign language technology research. *Universal Access in the Information Society*, pages 1–9.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous Measurement Scales in Human Evaluation of Machine Translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Thomas Hanke. 2004. Hamnosys - representing sign language data in language resources and language processing contexts. In *LREC 2004, Workshop proceedings : Representation and processing of sign languages*, pages 1–6. Paris : ELRA.

Thomas Hanke, Susanne König, Reiner Konrad, Gabriele Langer, Patricia Barbeito Rey-Geißler, Dolly Blanck, Stefan Goldschmidt, Ilona Hofmann, Sung-Eun Hong, Olga Jeziorski, Thimo Kleyboldt, Lutz König, Silke Matthes, Rie Nishio, Christian Rathmann, Uta Salden, Sven Wagner, and Satu Worseck. 2020. MEINE DGS. Öffentliches Korpus der Deutschen Gebärdensprache, 3. Release.

Thomas Hanke and Jakob Storz. 2008. ilex–a database tool for integrating sign language corpus linguistics and sign language lexicography. In *sign-lang@ LREC 2008*, pages 64–67. European Language Resources Association (ELRA).

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567*.

Mert Inan, Yang Zhong, Sabit Hassan, Lorna Quandt, and Malihe Alikhani. 2022. Modeling intensification for sign language generation: A computational approach. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2897–2911, Dublin, Ireland. Association for Computational Linguistics.

Zifan Jiang, Amit Moryossef, Mathias Müller, and Sarah Ebling. 2023a. Signsuisse dsgs/lsf/lis lexicon.

Zifan Jiang, Amit Moryossef, Mathias Müller, and Sarah Ebling. 2023b. Srf dsgs daily news broadcast: video and original subtitle data.

Trevor Johnston. 2010. From archive to corpus: Transcription and annotation in the creation of signed language corpora. *International Journal of Corpus Linguistics*, 15(1):106–131.

Trevor Johnston. 2011. Lexical Frequency in Sign Languages. *The Journal of Deaf Studies and Deaf Education*, 17(2):163–193.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Sang-Ki Ko, Chang Jo Kim, Hyedong Jung, and Choongsang Cho. 2019. Neural sign language translation based on human keypoint estimation. *Applied Sciences*, 9(13):2683.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Oscar Koller. 2020. Quantitative survey of the state of the art in sign language recognition. *arXiv preprint arXiv:2008.09918*.

Maria Kopf, Marc Schulder, and Thomas Hanke. 2021. Overview of Datasets for the Sign Languages of Europe.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

J R Landis and G G Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in Neural Machine Translation.

Lorraine Leeson. 2005. Making the effort in simultaneous interpreting. *Topics in Signed Language Interpreting: Theory and Practice, ed. by Terry Janzen.– Amsterdam.*

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1459–1469.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. 2019. MediaPipe: A Framework for Building Perception Pipelines. *CoRR*, abs/1906.08172.

Qingsong Ma, Yvette Graham, Timothy Baldwin, and Qun Liu. 2017. Further investigation into reference bias in monolingual evaluation of machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2466–2475, Copenhagen, Denmark. Association for Computational Linguistics.

Caio DD Monteiro, Christy Maria Mathew, Ricardo Gutierrez-Osuna, and Frank Shipman. 2016. Detecting and identifying sign languages through visual features. In *2016 IEEE International Symposium on Multimedia (ISM)*, pages 287–290. IEEE.

Sara Morrissey. 2011. Assessing three representation methods for sign language machine translation and evaluation. In *Proceedings of the 15th annual meeting of the European Association for Machine Translation (EAMT 2011), Leuven, Belgium*, pages 137–144.

Amit Moryossef and Yoav Goldberg. 2021. Sign Language Processing. `https://sign-language-processing.github.io/`.

Amit Moryossef and Mathias Müller. 2021. pose-format: Library for viewing, augmenting, and handling .pose files. `https://github.com/AmitMY/pose-format`.

Amit Moryossef, Mathias Müller, Anne Göhring, Zifan Jiang, Yoav Goldberg, and Sarah Ebling. 2023. An open-source gloss-based baseline for spoken to signed language translation. In *2nd International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*. Available at: `https://arxiv.org/abs/2305.17714`.

Amit Moryossef, Ioannis Tsochantaridis, Joe Dinn, Necati Cihan Camgöz, Richard Bowden, Tao Jiang, Annette Rios, Mathias Müller, and Sarah Ebling. 2021. Evaluating the immediate applicability of pose estimation for sign language recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 10166v1. 2021 ChaLearn Looking at People Sign Language Recognition in the Wild Workshop at CVPR.

Mathias Müller, Sarah Ebling, Eleftherios Avramidis, Alessia Battisti, Michèle Berger, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Cristina España-bonet, Roman Grundkiewicz, Zifan Jiang, Oscar Koller, Amit Moryossef, Regula Perrollaz, Sabine Reinhard, Annette Rios, Dimitar Shterionov, Sandra Sidler-miserez, and Katja Tissi. 2022. Findings of the first WMT shared task on sign language translation (WMT-SLT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 744–772, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023. Considerations for meaningful sign language machine translation based on glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 682–693, Toronto, Canada. Association for Computational Linguistics.

Ellen Ormel and Onno Crasborn. 2012. Prosodic correlates of sentences in signed languages: A literature review and suggestions for new types of studies. *Sign Language Studies*, 12(2):279–315.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Pamela Perniss, Asli Özyürek, and Gary Morgan. 2015. The influence of the visual modality on language structure and conventionalization: Insights from sign language and gesture. *Topics in Cognitive Science*, 7(1):2–11.

Elena Pizzuto and Paola Pietrandrea. 2001. The Notation of Signed Texts: Open Questions and Indications for Further Research. *Sign Language & Linguistics*, 4:29–45.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtskỳ. 2020. Transforming Machine Translation: a Deep Learning System Reaches News Translation Quality Comparable to Human Professionals. *Nature communications*, 11(1):1–15.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The Curious Case of Hallucinations in Neural Machine Translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Marcelo Sandoval-Castaneda, Yanhong Li, Bowen Shi, Diane Brentari, Karen Livescu, and Gregory Shakhnarovich. 2023. TTIC's Submission to WMT-SLT 23. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. 2020a. Adversarial training for multi-channel sign language production. In *The 31st British Machine Vision Virtual Conference*. British Machine Vision Association.

Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. 2020b. Everybody sign now: Translating spoken language to photo realistic sign language video. *arXiv preprint arXiv:2011.09846*.

Ben Saunders, Necati Cihan Camgöz, and Richard Bowden. 2020c. Progressive transformers for end-to-end sign language production. In *European Conference on Computer Vision*, pages 687–705.

Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. 2022. Signing at Scale: Learning to Co-Articulate Signs for Large-Scale Photo-Realistic Sign Language Production. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Rico Sennrich and Biao Zhang. 2019. Revisiting Low-Resource Neural Machine Translation: A Case Study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

Bowen Shi, Diane Brentari, Gregory Shakhnarovich, and Karen Livescu. 2022. Open-domain sign language translation learned from online video. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6365–6379, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anita Slonimska, Asli Özyürek, and Olga Capirci. 2021. Using Depiction for Efficient Communication in LIS (Italian Sign Language). *Language and Cognition*, 13(3):367–396.

Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. 2020. Text2Sign: Towards Sign Language Production Using Neural Machine Translation and Generative Adversarial Networks. *International Journal of Computer Vision*, 128(4):891–908.

Valerie Sutton. 1990. *Lessons in sign writing*. SignWriting.

Laia Tarrés, Gerard I. Gállego, Amanda Duarte, Jordi Torres, and Xavier Giró i Nieto. 2023. Sign language translation from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) :Workshops*.

David Uthus, Garrett Tanzer, and Manfred Georg. 2023. Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus.

Gül Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. 2021. Read and attend: Temporal localisation in sign language videos. In *CVPR*.

Gül Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. 2022. Scaling up sign spotting through sign language dictionaries. *International Journal of Computer Vision*, 130(6):1416–1439.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Carla Viegas, Mert Inan, Lorna Quandt, and Malihe Alikhani. 2023. Including facial expressions in contextual embeddings for sign language generation. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 1–10, Toronto, Canada. Association for Computational Linguistics.

Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *CoRR*, abs/1610.02424.

Harry Walsh, Ben Saunders, and Richard Bowden. 2022. Changing the representation: Examining language representation for neural sign language production. In *LREC 2022*.

Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. 2020. SOLOv2: Dynamic and Fast Instance Segmentation. In *Advances in Neural Information Processing Systems*, volume 33, pages 17721–17732. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Rosalee Wolfe, John C. McDonald, Thomas Hanke, Sarah Ebling, Davy Van Landuyt, Frankie Picron, Verena Krausneker, Eleni Efthimiou, Evita Fotinea, and Annelies Braffort. 2022. Sign language avatars: A question of representation. *Information*, 13(4):206.

Bencie Woll. 2013. 9091 The History of Sign Language Linguistics. In *The Oxford Handbook of the History of Linguistics*. Oxford University Press.

Baixuan Xu, Haochen Shi, Tianshi Zheng, Qing Zong, Weiqi Wang, Zhaowei Wang, and Yangqiu Song. 2023. KnowComp Submission for WMT23 Sign Language Translation Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including Signed Languages in Natural Language Processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.

Kayo Yin and Jesse Read. 2020. Better sign language translation with stmc-transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5975–5989.

A Details on shared task data and submission

A.1 Data resources

Direct download links: https://www.swissubase.ch/en/catalogue/studies/20452/19173/datasets/2327/2705/overview
Signsuisse lexicon (release 2.0): https://www.swissubase.ch/en/catalogue/studies/20452/19280/datasets/2350/2715/overview
SRF corpus poses and segmented subtitles (release 1.0): https://www.swissubase.ch/en/catalogue/studies/20452/19280/datasets/2343/2721/overview
Test sources as a tar ball (release 2.0): https://files.ifi.uzh.ch/cl/archiv/2023/easier/wmtslt/test_sources.v2.0.tar.gz
Test sources in WMT XML format for submissions: https://files.ifi.uzh.ch/cl/archiv/2023/easier/wmtslt/xml/

A.2 XML submission schema

```
<?xml version='1.0' encoding='utf-8'?>
<dataset id="slttest2022.de-dsgs">
  <doc origlang="de" id="srf.0">
    <src lang="de">
      <p>
        <seg id="0">Guten Abend meine Damen und Herren - willkommen zur "Tagesschau".</seg>
      </p>
    </src>
    <hyp system="YOUR SYSTEM NAME" language="dsgs">
      <p>
        <seg id="0">     https://www.your_hosting.com/your_url_for_this_segment
</seg>
      </p>
    </hyp>
  </doc>
</dataset>
```

B Appraise instructions to human evaluators

B.1 Sign-to-text direction

B.1.1 English

Below you see a document with 10 sentences in Swiss-German Sign Language (Deutschschweizer Gebärdensprache (DSGS)) (left columns) and their corresponding candidate translations in German (Deutsch) (right columns). Score each candidate sentence translation in the document context. You may revisit already scored sentences and update their scores at any time by clicking on a source video.

Assess the translation quality on a continuous scale using the quality levels described as follows:

- 0: Nonsense/No meaning preserved: Nearly all information is lost between the translation and source. Grammar is irrelevant.
- 2: Some Meaning Preserved: The translation preserves some of the meaning of the source but misses significant parts. The narrative is hard to follow due to fundamental errors. Grammar may be poor.
- 4: Most Meaning Preserved and Few Grammar Mistakes: The translation retains most of the meaning of the source. It may have some grammar mistakes or minor contextual inconsistencies.

- 6: Perfect Meaning and Grammar: The meaning of the translation is completely consistent with the source and the surrounding context. The grammar is also correct.

Please score the overall document translation quality (you can score the whole document only after scoring all individual sentences first). Assess the translation quality on a continuous scale using the quality levels described as follows:

- 0: Nonsense/No meaning preserved: Nearly all information is lost between the translation and source. Grammar is irrelevant.
- 2: Some Meaning Preserved: The translation preserves some of the meaning of the source but misses significant parts. The narrative is hard to follow due to fundamental errors. Grammar may be poor.
- 4: Most Meaning Preserved and Few Grammar Mistakes: The translation retains most of the meaning of the source. It may have some grammar mistakes or minor contextual inconsistencies.
- 6: Perfect Meaning and Grammar: The meaning of the translation is completely consistent with the source and the surrounding context. The grammar is also correct.

B.1.2 German

Unten sehen Sie ein Dokument mit 10 Sätzen in Deutschschweizer Gebärdensprache (DSGS) (linke Spalten) und die entsprechenden möglichen Übersetzungen auf Deutsch (rechte Spalten). Bewerten Sie jede mögliche Übersetzung des Satzes im Kontext des Dokuments. Sie können bereits bewertete Sätze jederzeit durch Anklicken eines Eingabevideos erneut aufrufen und die Bewertung aktualisieren.

Bewerten Sie die Übersetzungsqualität auf einer kontinuierlichen Skala mit Hilfe der nachfolgend beschriebenen Qualitätsstufen:

- 0: Unsinn/Bedeutung nicht erhalten: Fast alle Informationen zwischen Übersetzung und Eingabevideo sind verloren gegangen. Die Grammatik ist irrelevant.
- 2: Ein Teil der Bedeutung ist erhalten: Die Übersetzung behält einen Teil der Bedeutung der Quelle bei, lässt aber wichtige Teile aus. Die Erzählung ist aufgrund von grundlegenden Fehlern schwer zu verstehen. Die Grammatik kann mangelhaft sein.
- 4: Der grösste Teil der Bedeutung ist erhalten und es gibt nur wenige Grammatikfehler: Die Übersetzung behält den grössten Teil der Bedeutung der Quelle bei. Sie kann einige Grammatikfehler oder kleinere kontextuelle Unstimmigkeiten aufweisen.
- 6: Perfekte Bedeutung und Grammatik: Die Bedeutung der Übersetzung stimmt vollständig mit der Quelle und dem umgebenden Kontext (falls zutreffend) überein. Auch die Grammatik ist korrekt.

Bitte bewerten Sie die Übersetzungsqualität des gesamten Dokuments. (Sie können das Dokument erst bewerten, nachdem Sie zuvor alle Sätze einzeln bewertet haben.) Bewerten Sie die Übersetzungsqualität auf einer kontinuierlichen Skala mit Hilfe der nachfolgend beschriebenen Qualitätsstufen:

- 0: Unsinn/Bedeutung nicht erhalten: Fast alle Informationen zwischen Übersetzung und Eingabevideo sind verloren gegangen. Die Grammatik ist irrelevant.
- 2: Ein Teil der Bedeutung ist erhalten: Die Übersetzung behält einen Teil der Bedeutung der Quelle bei, lässt aber wichtige Teile aus. Die Erzählung ist aufgrund von grundlegenden Fehlern schwer zu verstehen. Die Grammatik kann mangelhaft sein.
- 4: Der grösste Teil der Bedeutung ist erhalten und es gibt nur wenige Grammatikfehler: Die Übersetzung behält den grössten Teil der Bedeutung der Quelle bei. Sie kann einige Grammatikfehler oder kleinere kontextuelle Unstimmigkeiten aufweisen.
- 6: Perfekte Bedeutung und Grammatik: Die Bedeutung der Übersetzung stimmt vollständig mit der Quelle und dem umgebenden Kontext (falls zutreffend) überein. Auch die Grammatik ist korrekt.

C Feedback from evaluators

Tables 9 and 10 detail for each evaluator the feedback answers and comments regarding the human evaluation procedure and the Appraise system. All three evaluators submitted a response.

	Answer 1	**Answer 2**	**Answer 3**
What is your experience in assessing machine translation outputs?			
	Low: I have done it once or a long time ago	Moderate: I have done it a few times	Low: I have done it once or twice before, or a long time ago
Please specify how much you agree or disagree with the following statements.			
Generally, my experience with the tool was positive	Agree	Agree	Agree
Instructions were clear	Neutral	Strongly agree	Strongly agree
Quality levels 0-6 were helpful to me	Neutral	Neutral	Agree
Source videos were understandable	Strongly agree	Agree	Strongly Agree
There was too much repetitiveness	Strongly agree	Neutral	Strongly agree
Documents were too long	Disagree	Disagree	Neutral
Segments were too short	Disagree	Disagree	Disagree
In some cases, the context was insufficient	Neutral	Neutral	Disagree
I experienced technical issues	Neutral	Neutral	Disagree
I would be willing to do similar work in the future	Agree	Agree	Agree
This evaluation campaign featured the Direct Assessment with Scalar Quality Metrics method. What do you think about this method? On a scale between -3 (negative) and 3 (positive) it was...			
difficult/easy	+1	+3	+3
stressful/relaxed	0	+3	+2
laborious/effortless	+2	+2	-2
slow/fast	+2	+2	0
inefficient/efficient	+2	+2	+2
boring/exciting	-1	+2	0
complicated/simple	+1	+2	+3
annoying/enjoyable	-1	+2	0
limiting/creative	-1	0	0
impractical/practical	0	+2	+3

Table 9: Feedback from evaluators about the human evaluation setup and the Appraise platform.

Answer 1		Answer 2	Answer 3
Please provide more details related to the statements above that you think can be useful to us. What was most troublesome? What could we improve?			
(original in German) - Ich hätte ein grösseres Video geschätzt (ohne dass ich das jedes Mal aktiv anklicken muss) > Z.B. bei Klicken auf Play, automatische Vergrösserung und bei Ende der Wiedergabe automatisch zurück auf die Skala. - Die Videoschnitte waren - v.a. bei einem Modell (langer Lag!) - sehr schlecht. Video und Text stimmten deshalb oft nicht überein. Schwierig für die Beurteilung! - Es kam oft vor, dass ganze Dokumente schon auf einen Blick als "komplett falsch" ersichtlich waren (Texte komplett unverständlich). Da wäre es hilfreich, wenn man ein gesamtes Dokument als "ROT" beurteilen könnte, ohne jedes einzelne Video zu beurteilen.	(translated into English) - I would have appreciated a larger video (without having to actively click that every time) > E.g. when clicking play, automatic enlargement and at the end of playback automatically back to the scale. - The video cuts were - especially with one model (long lag!) - very bad. Video and text therefore often did not match. Difficult for the evaluation! - It often happened that whole documents appeared at a glance as "completely wrong" (texts completely incomprehensible). There it would be helpful if one could judge a whole document as "RED" without judging every single video.	Some of the film clips were poorly edited and therefore did not match the translated text. Certain written formulations are not common in Switzerland. There are some very German formulations. The German text was taken over, there was no real translation.	The large amount of nonsense translations could lead to the fact that one does not work concentrated any more.
What were the main or most common issues with the automatic translations?			
(original in German) Es gab wenig Probleme technischer Art. Nur 1x kein Zugang zum Dokument. Ab und zu (aber selten!) eine Meldung, dass die "Resultate" nicht angenommen/gespeichert werden konnten.	(translated into English) There were few problems of a technical nature. Only 1x no access to the document. Now and then (but rarely!) a message that the "results" could not be accepted/saved.	Some of the film clips were poorly edited and therefore did not match the translated text.	The large amount of nonsense translations.

Table 10: Feedback comments from evaluators about the human evaluation setup and the Appraise platform.

Findings of the WMT 2023 Shared Task on Parallel Data Curation

Steve Sloto
Microsoft
ssloto@microsoft.com

Brian Thompson
AWS AI Labs
brianjt@amazon.com

Huda Khayrallah
Microsoft
hkhayrallah@microsoft.com

Tobias Domhan
Amazon
domhant@amazon.de

Thamme Gowda
Microsoft
thammegowda@microsoft.com

Philipp Koehn
Johns Hopkins University
phi@jhu.edu

Abstract

Building upon prior WMT shared tasks in document alignment and sentence filtering, we posed the open-ended shared task of finding the best subset of possible training data from a collection of Estonian-Lithuanian web data. Participants could focus on any portion of the end-to-end data curation pipeline, including alignment and filtering. We evaluated results based on downstream machine translation quality. We release processed Common Crawl data, along with various intermediate states from a strong baseline system, which we believe will enable future research on this topic.

1 Introduction

A machine translation (MT) system is only as good as the data it is trained on. However, the academic research community often overlooks the details of this task, using pre-curated corpora.

To promote research in this area, this shared task[1] focuses on finding pairs of sentences or documents that are translations of each other based on a collection of web crawled data. MT models are trained by the organizers on the data found by participants, and performance is then judged using automatic metrics. This shared task builds on prior shared tasks on document alignment (Buck and Koehn, 2016a) and sentence filtering (Koehn et al., 2018, 2019, 2020). However, this task is intentionally open-ended, and designed to allow participants to improve on various different parts of the data curation pipeline.

We chose the Estonian-Lithuanian language pair for several reasons. The amount of data we extracted in that language pair was enough to train a reasonable MT model, while being small enough that the task was still accessible to academic participants with limited hardware resources. We avoided English, as many toolkits are developed/optimized on English data, and results on English may not generalize well. And finally, we avoided languages which were closely related, as this could favor methods which do not generalize well.

To lower the barrier to entry and allow participants to focus their research and compute resources, we release intermediate stages of a strong baseline data curation system. We encourage future work to build upon resources provided in this shared task.

This paper gives an overview of the task, presents its results, and provides some analysis.

2 Related work

Parallel data has been required for training machine translation systems ever since the field transitioned to statistical machine translation (Brown et al., 1990). To train that first statistical system, Brown et al. aligned English-French sentences from the proceedings of the Canadian Parliament, often referred to as Hansards, using a very simple system to segment each side into sentences and then align them using only sentence length (Brown et al., 1991). The field of parallel data curation has come a long way since then, with modern methods extracting billions of sentence pairs in hundreds of languages, as opposed to the few million enabled by Hansards.

Currently, there are two main approaches to parallel data curation: (1) document and sentence alignment, and (2) comparable corpora methods.

Document & Sentence alignment The first approach is very similar in spirit to that used on Handards: Parallel documents are identified and then document pairs are aligned at the sentence level to produce sentence-level translation pairs. These steps are referred to as document alignment and sentence alignment, respectively. The web has become the default source of documents (Resnik, 1998), where businesses, governments, and individuals regularly release documents and translations of

[1] http://www2.statmt.org/wmt23/data-task.html

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 95–102

those documents–for example a user manual that is published in several languages. A very simple and computationally inexpensive approach to finding parallel documents is to locate URLs which differ in no more than a language code (Resnik and Smith, 2003). However, more accurate (and computationally expensive) methods have also been developed which look for documents which appear to contain similar information, for example by translating all documents into one language and then finding pairs via TF-IDF similarity (Buck and Koehn, 2016b). More recent approaches to document alignment have relied on finding similar vectors after converting documents into multilingual vectors, created via combining sentence embeddings (Thompson and Koehn, 2020) or by embedding entire documents (Guo et al., 2019). A WMT shared task on document alignment was held in 2016 (Buck and Koehn, 2016a).

Once parallel documents have been located, they are sentence aligned. Sentence alignment consists of finding a bipartite graph which matches minimal groups of sentences that are translations of each other. This is necessary because content may have been inserted or deleted in the translation process, and sentences may have been combined or split in the translation process. Additionally, sentence segmentation errors may cause sentences to be split or combined. An example of an early sentence alignment algorithm is Gale-Church (Gale and Church, 1993), which like the original IBM system uses only the length of each sentence, making it very computationally efficient but not particularly accurate. Bleualign (Sennrich and Volk, 2010, 2011) used an MT system to convert one text into the language of the other and then performed n-gram matching, similar to the BLEU MT metric (Papineni et al., 2002). A more recent sentence aligner is Vecalign (Thompson and Koehn, 2019), which uses multilingual sentence embeddings and a dynamic programming approximation (Salvador and Chan, 2007) which makes the algorithm linear with respect to the number of sentences being aligned. Widely used datasets created via document and sentence alignment include Paracrawl (Bañón et al., 2020) and CCAlign (El-Kishky et al., 2020).

Comparable Corpora A recent alternative to document and sentence alignment is to discard document information and simply create a collection of sentences in each language, and then find translation pairs by looking for sentences which are nearby by in a multilingual embedding space. LASER (Artetxe and Schwenk, 2019) was proposed for this task. The authors additionally proposed a margin-based score which gives preference to sentence pairs which are more similar to one another than other potential matches by at least a minimum margin. Approximate nearest neighbor search (Johnson et al., 2019) is used to make the search for sentence pairs tractable. Examples of widely-used datasets created via the comparable corpora method include Wikimatrix (Schwenk et al., 2021a) and CCMatrix (Schwenk et al., 2021b).

2.1 Parallel Corpora Filtering

Once data has been aligned, it is customary—especially for data coming from the web—to perform data filtering to remove low quality translation pairs before using the data for training, as unfiltered web-crawled data harms translation performance (Khayrallah and Koehn, 2018). There have been three prior shared tasks on bitext filtering at WMT (Koehn et al., 2018, 2019, 2020).

Popular approaches to data filtering include LASER margin filtering (Chaudhary et al., 2019), using an approach similar to the comparable corpora method described above, and dual conditional cross entropy (Junczys-Dowmunt, 2018), which trains NMT models on held-out clean data in both the forward and reverse directions and uses them to compute cross-entropy scores for the data being filtered. Sentence pairs with divergent or poor cross-entropies are down-weighted.

3 Shared Task Definition

This shared task presented the open-ended problem of finding the best possible subset of aligned sentence pairs from unaligned documents sourced from the internet. Participants were evaluated on downstream machine translation system performance.

Parallel data curation from web can be computationally demanding due to the sheer scale of webcrawled data. For this reason, in addition to our documents, we also released pre-computed intermediate steps from a baseline, so participants can choose to focus on one aspect of the task (e.g. sentence filtering.)

For this shared task, the organizers provided:

- Web-crawled data, as unique sentences or unique documents

- LASER2 sentence embeddings
- K-nearest neighbors by cosine similarity from our baseline
- End-to-end scripts for MT training and evaluation

End-to-end scripts enabled participants to supply a set of sentence ids and train and evaluate a Sockeye MT model (Hieber et al., 2022). Alongside the scripts, we provided a simple baseline based on 1-best cosine similarity.

Participants were allowed to use only pre-trained models and datasets publicly released with a research-friendly license on or before May 1, 2023.

3.1 Dataset

All of our inputs were derived from the 2023-06 snapshot of Common Crawl. We extracted the plain text from HTML using the *trafilatura* library (Barbaresi, 2021), and ran the first 2,000 characters through the 176-language fasttext language id model (Joulin et al., 2016a,b).

We kept all documents classified as Estonian or Lithuanian, unless their hostnames were included in the following lists from the blocklist project:[2] abuse, basic, crypto, drugs, fraud, gambling, malware, phishing, piracy, porn, ransomware, redirect, scam, torrent. No further data filtering was performed.

We split documents into paragraphs at line breaks, and segmented resulting paragraphs into sentences using the Media Cloud sentence splitter.[3]

Each unique sentence was given a Globally Unique IDentifier (GUID) and tagged with a language id based on *fastText*.

3.1.1 Dataset Statistics

Our dataset includes documents taken from 402,920 hosts. Only 24,319 of these hosts included documents in both languages. Table 1 includes overall counts on a per language basis.

3.1.2 Intermediate Outputs From Baselines

We provide participants with intermediate outputs from our baseline systems as additional resources, such that prospective participants could be able to access sentence embedding or sentence pair similarity information without needing computational resources to create these themselves.

[2] https://github.com/blocklistproject/Lists

[3] https://github.com/mediacloud/sentence-splitter

	Estonian	Lithuanian
# Hosts	199,813	227,426
# Documents	3,449,211	4,571,947
# Sentences	53,234,425	63,488,253
# Sents w/ LangId	36,870,945	46,969,824

Table 1: Counts of unique hosts, documents, sentences, and sentences identified as the correct language in our dataset

We provide outputs of embedding each sentence with the LASER 2 model (Heffernan et al., 2022). We also release a smaller version of the embeddings, projected down to 128 dimensions via PCA and converted to float16.

To create baseline sentence pair alignments, we removed sentences detected as non-Estonian or non-Lithuanian, and used the FAISS library (Johnson et al., 2019) to index our LASER2 embeddings for fast retrieval. We applied L2 normalization to the embeddings, and added them to a flat inner product index, so that the resulting scores were equivalent to cosine similarity. We queried each index with embeddings in the other language, and returned the top eight results. These raw cosine similarity scores are shared with participants as a potential resource, and serve as the basis for our baseline submissions.

4 Evaluation

We evaluated submissions by using the curated data to train machine translation systems.

For preprocessing, we split sentences into subwords by applying Byte-Pair Encoding (BPE) (Sennrich et al., 2016) using 32,000 merge operations. The BPE vocabulary is learned jointly for the source and target language. We apply a minimum vocabulary frequency of 100 per language.

We use Sockeye (Hieber et al., 2022) to train Transformer (Vaswani et al., 2017) translation models with 512 hidden units, 8 attention heads, 6 layers and feed-forward layers of size 2048. For training we use an effective batch size of 400k target tokens. We use 4096 target tokens per GPU, and gradient accumulation to obtain 400k target tokens regardless of the number of GPUs.

We use the Adam optimizer (Kingma and Ba, 2014) with $\beta_1 = 0.9$ and $\beta_1 = 0.98$, an initial learning rate of 0.06325, a linear warmup for 4000 updates and an inverse square root learning rate

# Sents	Min Margin Score	EMEA	EUbookshop	Europarl	JRC-Acquis	average
1.6M	1.048	21.1	23.2	20.3	17.9	20.6
3.2M	1.027	**21.9**	**23.6**	**20.8**	**18.5**	**21.2**
4.8M	1.019	21.7	**23.6**	**20.8**	18.4	21.1
6.4M	1.013	21.6	23.4	**20.8**	18.3	21.0
8.0M	0.900	21.3	23.3	20.6	18.1	20.8

Table 2: Comparison of different training data sizes and margin score cutoffs on development set BLEU.

decay. Checkpoints are written every 500 updates and training is stopped once validation perplexity does not improve for 12 checkpoints. The checkpoint with the lowest validation perplexity is used as the final checkpoint.

All systems are trained on nodes with 8 V100 GPUs. We use BLEU (Papineni et al., 2002) and chrF (Popović, 2015) as quality metrics. Evaluation metrics are computed using Sacrebleu (Post, 2018).

We considered data from four domains for evaluation: EMEA,[4] EUbookshop,[5] Europarl,[6] JRC-Acquis,[7] and EUconst.[8] All data is released by OPUS (Tiedemann, 2012). From each domain, we created a dev, test, and held-out-test set. We use up to 10,000 lines for each. If less data is available, it is split between the three sets. We also kept EUconst as a held-out domain.

5 Systems

We report the results of four different systems: the baseline, two participant systems, and a contrastive system.

5.1 Baseline

The naive baseline was designed to give participants a simple end-to-end system, so they could focus on any part of the pipeline to improve upon. While participants were not required to build upon the baseline, doing so lowered the barrier to entry.

As described in Section 3.1.2, we used the LASER 2 model to embed all Estonian and Lithuanian sentences, indexed them with FAISS, and computed the eight nearest neighbors' cosine similarities for each sentence in each language. We provided these cosine similarity scores as an additional resource for participants.

Our naive baseline was created by taking all sentence pairs whose cosine similarities whose 1-best neighbor exceeded or matched the threshold of 0.9 in the Estonian → Lithuanian direction, meaning that multiple target sentences could be aligned to the same source.

This naive baseline was designed to be an end-to-end solution to allow participants to improve on any of the individual parts (filtering, alignment, margin scoring, etc).

5.2 Steingrímsson

Steingrímsson (2023b) first perform document alignment and sentence alignment, and then use matches from the provided top1-cosine data for sentences which were not aligned via document/sentence alignment.

They perform sentence alignment of all document pairs within each web domain and score the alignments to locate document pairs, similar to Thompson and Koehn (2020), to find high-quality document pairs. They use the recently proposed SentAlign[9] (Steingrímsson, 2023a; Steingrímsson et al., 2023b) sentence aligner, which in turn uses LaBSE (Feng et al., 2022) sentence embeddings.

They also perform extensive bitext filtering, using several different language ID tools and the filtering method proposed in Steingrímsson et al. (2023a) which uses perplexities of a GPT-2 model (Radford et al., 2019), LAESR embeddings (Chaudhary et al., 2019), NMTScore (Vamvas and Sennrich, 2022) using Prism (Thompson and Post, 2020a,b), and WAScore (Steingrímsson et al., 2021), as well as Bicleaner AI (Zaragoza-Bernabeu et al., 2022).

5.3 Nguyen-Hoang et al.

Nguyen-Hoang et al. (2023) focus on using the phrase based dictionary to distill the high-quality sentences and making a pipeline to re-ranking the top-K cosine similarity.

They begin with the released data, and an MGiza-based (Gao and Vogel, 2008) dictionary. They then extract sentence pairs using the a top-1 cosine score

[4] https://opus.nlpl.eu/EMEA.php
[5] https://opus.nlpl.eu/EUbookshop.php
[6] https://opus.nlpl.eu/Europarl.php
[7] https://opus.nlpl.eu/JRC-Acquis.php
[8] https://opus.nlpl.eu/EUconst.php
[9] https://github.com/steinst/SentAlign

	BLEU				ChrF			
Test	EMEA	EUbooks	Europarl	JRC-Acquis	EMEA	EUbooks	Europarl	JRC-Acquis
Top1_cosine	18.1	20.1	18.4	25.7	49.4	**53.0**	52.1	**55.7**
Nguyen-Hoang et al.	18.5	**20.4**	**19.1**	**25.8**	48.9	52.5	**52.5**	55.5
Steingrímsson	**20.4**	20.2	18.7	25.4	**51.4**	52.8	52.0	54.9
MarginScore 3.2M	21.5	22.4	20.2	27.9	52.5	54.7	53.4	57.8

Table 3: Test set BLEU and ChrF scores. Top1_cosine is the baseline, and Marginscore 3.2M is the contrastive system.

	BLEU					ChrF				
Held-out	EMEA	EUbooks	Europarl	JRC-A	EUconst	EMEA	EUbooks	Europarl	JRC-A	EUconst
Top1_cosine	18.7	14.0	18.2	22.9	23.8	49.8	**47.6**	52.4	54.0	**58.5**
Nguyen-Hoang et al.	19.3	**14.4**	**19.1**	**23.5**	**25.1**	49.7	47.4	**52.9**	**54.2**	58.3
Steingrímsson	**21.0**	**14.5**	18.7	23.1	23.2	**52.1**	**47.6**	52.3	53.6	57.8
MarginScore 3.2M	21.9	16.1	20.5	25.4	27.6	52.9	48.9	53.8	56.2	60.9

Table 4: Held-out test BLEU and ChrF scores. Top1_cosine is the baseline, and Marginscore 3.2M is the contrastive system.

and a threshold. From there, the dictionary is used to translate the source sentences. These dictionary-translated sentences are then compared with the translation from the baseline data. The translation from the baseline data is filtered based on the edit distance. Then a NMT model is trained, and the final threshold is set based on NMT model performance.

Nguyen-Hoang et al. (2023) also perform an analysis on the cosine score threshold, demonstrating how varying this value impacts both corpus size and translation quality.

5.4 Contrastive System

The participants in this task both performed data filtering on top of the the top-1 cosine baseline.

Since no participants experimented with using margin scoring, which Schwenk et al. (2021b) found significant for improving the quality of LASER-based mining, the organizers created a stronger contrastive system that did so.

We calculated margin scores for our four nearest neighbors in both directions. We performed competitive linking,[10] such that each sentence appeared only once in our contrastive submission. Although we computed cosine similarities for the eight nearest neighbors, no appreciable difference was found in MT quality by using k=8 instead of k=4 when computing margin scores.

We sorted our data by margin score and compared different data sizes, as shown in Table 2. We used a minimum margin score of 1.027 and data size of 3.2 million lines since it scored the highest on all development sets and had the highest average score.

6 Results

Table 3 and Table 4 show the BLEU and ChrF results of the naive top-1 cosine baseline, participant submissions, and the contrastive margin score system. Of the baseline and two participant systems, we bold the best and systems within 0.1 of the best. Overall, both participants improved over the naive baseline. On the held-out test sets, Steingrímsson had higher BLEU on EMEA and EUbookshop, while Nguyen-Hoang et al. had higher BLEU on Europarl, JRC-Acquis, and the held-out domain of EUconst.

We see that the contrastive margin score system outperforms the naive top-1 cosine baseline. This confirms the finding of Schwenk et al. (2021b) that margin scoring outperforms raw cosine similarity. The contrastive margin score system also outperforms the participant submissions that directly build and improve upon the naive top-1 cosine baseline.

Data filtering and alignment tend to be complimentary, so the filtering methods proposed by the

[10] Referred to as the "max strategy" by Schwenk et al. (2021b).

participants would likely improve upon the contrastive margin score system if they were applied on top of it.

7 Conclusion

While data curation is the first step in the training of any MT (or machine learning) model, this tends to be a less-published-upon topic in academic research.

In this shared task, we have released the processed webcrawled data, and a baseline system with intermediate outputs. We hope this task lowers the barrier of entry and allow participants to focus on any aspect of the data curation pipeline (document alignment, sentence alignment, filtering, etc.) We have trained and evaluated MT systems on the datasets curated by participating teams. We have presented results for two participant submissions, in addition to two more systems built by the shared task organizers.

We hope this work serves as a building block for future research on this topic.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Adrien Barbaresi. 2021. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131, Online. Association for Computational Linguistics.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, John Lafferty, Robert L Mercer, and Paul S Roossin. 1990. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Peter F Brown, Jennifer C Lai, and Robert L Mercer. 1991. Aligning sentences in parallel corpora. In *29th Annual Meeting of the Association for Computational Linguistics*, pages 169–176.

Christian Buck and Philipp Koehn. 2016a. Findings of the WMT 2016 bilingual document alignment shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 554–563, Berlin, Germany. Association for Computational Linguistics.

Christian Buck and Philipp Koehn. 2016b. Quick and reliable document alignment via TF/IDF-weighted cosine distance. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 672–678, Berlin, Germany. Association for Computational Linguistics.

Vishrav Chaudhary, Yuqing Tang, Francisco Guzmán, Holger Schwenk, and Philipp Koehn. 2019. Low-resource corpus filtering using multilingual sentence embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 261–266, Florence, Italy. Association for Computational Linguistics.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio. Association for Computational Linguistics.

Mandy Guo, Yinfei Yang, Keith Stevens, Daniel Cer, Heming Ge, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Hierarchical document encoder for parallel corpus mining. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 64–72, Florence, Italy. Association for Computational Linguistics.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Felix Hieber, Michael Denkowski, Tobias Domhan, Barbara Darques Barros, Celina Dong Ye, Xing Niu, Cuong Hoang, Ke Tran, Benjamin Hsu, Maria Nadejde, et al. 2022. Sockeye 3: Fast neural machine translation with pytorch. *arXiv preprint arXiv:2207.05851*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742, Online. Association for Computational Linguistics.

Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino. 2019. Findings of the WMT 2019 shared task on parallel corpus filtering for low-resource conditions. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 54–72, Florence, Italy. Association for Computational Linguistics.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the WMT 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 726–739, Belgium, Brussels. Association for Computational Linguistics.

Minh-Cong Nguyen-Hoang, Van Vinh Nguyen, and Le-Minh Nguyen. 2023. A fast method to filter noisy parallel data WMT2023 shared task on parallel data curation. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Philip Resnik. 1998. Parallel strands: a preliminary investigation into mining the web for bilingual text. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 72–82, Langhorne, PA, USA. Springer.

Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Stan Salvador and Philip Chan. 2007. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021a. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021b. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich and Martin Volk. 2010. MT-based sentence alignment for OCR-generated parallel texts. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*, Denver, Colorado, USA. Association for Machine Translation in the Americas.

Rico Sennrich and Martin Volk. 2011. Iterative, MT-based sentence alignment of parallel texts. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 175–182, Riga, Latvia. Northern European Association for Language Technology (NEALT).

Steinþór Steingrímsson. 2023a. *Effectively compiling parallel corpora for machine translation in resource-scarce conditions*. Ph.D. thesis, Reykjavik University.

Steinþór Steingrímsson. 2023b. A sentence alignment approach to document alignment and multi-faceted filtering for curating parallel sentence pairs from web-crawled data. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore. Association for Computational Linguistics.

Steinþór Steingrímsson, Hrafn Loftsson, and Andy Way. 2023a. Filtering matters: Experiments in filtering training sets for machine translation. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 588–600, Tórshavn, Faroe Islands. University of Tartu Library.

Steinþór Steingrímsson, Hrafn Loftsson, and Andy Way. 2023b. Sentalign: Accurate and scalable sentence alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Singapore, Singapore. Association for Computational Linguistics.

Steinþór Steingrímsson, Pintu Lohar, Hrafn Loftsson, and Andy Way. 2021. Effective bitext extraction from comparable corpora using a combination of three different approaches. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, pages 8–17, Online (Virtual Mode). INCOMA Ltd.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Brian Thompson and Philipp Koehn. 2020. Exploiting sentence order in document alignment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5997–6007, Online. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020a. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020b. Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the Fifth Conference on Machine Translation*, pages 561–570, Online. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Jannis Vamvas and Rico Sennrich. 2022. NMTScore: A multilingual analysis of translation-based text similarity measures. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 198–213, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jaume Zaragoza-Bernabeu, Gema Ramírez-Sánchez, Marta Bañón, and Sergio Ortiz Rojas. 2022. Bicleaner AI: Bicleaner goes neural. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 824–831, Marseille, France. European Language Resources Association.

Samsung R&D Institute Philippines at WMT 2023

Jan Christian Blaise Cruz
Samsung R&D Institute Philippines
jcb.cruz@samsung.com

Abstract

In this paper, we describe the constrained MT systems submitted by Samsung R&D Institute Philippines to the WMT 2023 General Translation Task for two directions: en→he and he→en. Our systems comprise of Transformer-based sequence-to-sequence models that are trained with a mix of best practices: comprehensive data preprocessing pipelines, synthetic backtranslated data, and the use of noisy channel reranking during online decoding. Our models perform comparably to, and sometimes outperform, strong baseline unconstrained systems such as **mBART50 M2M** and **NLLB 200 MoE** despite having significantly fewer parameters on two public benchmarks: FLORES-200 and NTREX-128.

1 Introduction

This paper describes Samsung R&D Institute Philippines's submission to the WMT 2023 General Translation task. We participate in two translation directions: en→he and he→en, submitting two **constrained** single-direction models based on the Transformer (Vaswani et al., 2017) sequence-to-sequence architecture. We employ a number of best practices, using a comprehensive data preprocessing pipeline to ensure parallel data quality, create synthetic data through carefully-curated backtranslation, and use reranking methods to select the best candidate translations.

Our systems achieve strong performance on public benchmarks: 44.24 BLEU and 33.77 BLEU for FLORES-200 and NTREX-128 en→he, respectively, and; 42.42 BLEU and 36.89 BLEU on FLORES-200 and NTREX-128 he→en, respectively. Our systems outperform **mBART50 M2M** and slightly underperform against **NLLB 200 MoE** despite having significantly less parameters compared to these unconstrained baselines.

We detail our data preprocessing, model training, data augmentation, and translation methodology. Additionally, we illustrate hyperparameter sweeping setups and study the effects of hyperparameters during online decoding with reranking.

2 Methodology

2.1 Data Preprocessing

Given that a significant portion of the training dataset is synthetically-aligned, we need to use a comprehensive data preprocessing pipeline to ensure good translation quality. In particular, we use a combination of heuristic-based, ratio-based, and embedding-based methods to filter our data.

Heuristic-based The following heuristic-based filters based on Cruz and Cheng (2021) are used before applying the others:

- **Language Filter** – We use use `pycld3`[1] to filter out sentence pairs where one or both sentences have more than 30% tokens that are neither English nor Hebrew.
- **Named Entity Filter** – We use NER models (Bareket and Tsarfaty, 2021; Yang and Zhang, 2018) to check if both sentences in a pair have matching entities (if any). Pairs that contain entities that do not match are removed.
- **Numerical Filter** – If one sentence in a pair has a number (ordinal, date, etc.), we also check the other sentence if a matching number is present. If a match is not detected, the pair is removed.

Ratio-based We employ ratio-based filters on tokenized sentence pairs following Cruz and Sutawika (2022) and Sutawika and Cruz (2021). We first tokenize using SacreMoses[2] then apply the following ratio-based filters:

[1] https://pypi.org/project/pycld2/
[2] https://github.com/alvations/sacremoses

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 103–109
December 6–7, 2023.

	Pairs	Words (en)	Words (he)
Original	72,459,348	701,991,594	566,555,530
Original Filtered	48,278,395	385,975,984	312,639,617
Synthetic en→he	10,000,000	165,595,289	145,849,940
Synthetic en→he Filtered	7,143,725	115,239,312	95,954,020
Synthetic he→en	73,278,018	1,471,827,973	1,056,677,671
Synthetic he→en Filtered	47,372,416	659,409,236	541,376,459

Table 1: Corpus Statistics. "Filtered" refers to the number of pairs / words that remain after the filtering script is applied to the dataset. Note that "Words" is an approximation gathered by using the `wc -l *` command on the plaintext files.

- **Length Filter** – We remove pairs containing sentences with more than 140 characters.
- **Token Length Filter** – We remove pairs that contain sentences with tokens that are more than 40 characters long.
- **Character to Token Ratio** – We remove pairs where the ratio between character count and token count in at least one sentence is greater than 12.
- **Pair Token Ratio** – We remove pairs where the ratio of tokens between the source and target sentences is greater than 4.
- **Pair Length Ratio** – We remove pairs where the ratio between the string lengths of the source and target sentences is greater than 6.

Embedding-based Finally, we experiment with the use of sentence embedding models to compute embedding-based similarity between a sentence pair. We use LaBSE (Feng et al., 2020) models to embed both the source and target sentences then compute a cosine similarity score between the two. The pair must have a similarity score $0.7 \leq s \leq 0.96$ to be kept.

Statistics on the original and filtered corpus can be found on Table 1.

After preprocessing the parallel data, we learn a shared BPE (Sennrich et al., 2015b) vocabulary using SentencePiece[3] (Kudo and Richardson, 2018) with 32,000 units. All models in this paper use the same shared vocabulary.

2.2 Model Architecture

We experiment with two model sizes for each language pair: a **`Base`** model with 65M parameters and a **`Large`** model with 200M parameters. Both models use the standard Transformer (Vaswani et al., 2017) sequence-to-sequence architecture and are trained using Fairseq (Ott et al., 2019) with the hyperparameters listed in Table 2.

Training Hyperparameters	
Parameters	65M and 200M
Vocab Size	32,000
Tied Weights	Yes
Dropout	0.3
Attention Dropout	0.1
Weight Decay	0.0
Label Smoothing	0.1
Optimizer	Adam
Adam Betas	β_1=0.90, β_2=0.98
Adam ϵ	ϵ=1e-6
Learning Rate	7e-4
Warmup Steps	4,000
Total Steps	1,000,000
Batch size	64,000 tokens

Table 2: Hyperparameters used during training. When reporting model sizes, **`Base`** refers to 65M parameters, while **`Large`** refers to 200M.

We parallelize with 8 NVIDIA Tesla P100 GPUs and initially train for a total of 100K steps for experimentation. For the submitted systems trained with backtranslated data, we train for a total of 1M steps.

2.3 Backtranslation

We use backtranslation (Sennrich et al., 2015a) as a form of data augmentation to improve our initial models. We generate synthetic data via combined top-k and nucleus sampling:

$$\sum_{i=0}^{\delta_k} P(\hat{y}_i^{(T)}|x; \hat{y}^{(T-1)}) * \delta_{temp} \leq \delta_p \quad (1)$$

[3] https://github.com/google/sentencepiece

Backtranslation Hyperparameters	
Top-k (δ_k)	50
Top-p (δ_p)	0.93
Temperature (δ_t)	0.7
Beam	1.0
Length Penalty	1.0

Table 3: Hyperparameters used during backtranslation.

where δ_k is the top values considered for top-k sampling, δ_{temp} is the temperature hyperparameter, and δ_p is the maximum total probability for nucleus sampling.

Backtranslation is only performed once using the provided monolingual data. We produce a total of 10,000,000 synthetic sentences for the en→he direction and 73,278,018 synthetic sentences for the he→en direction. The same data preprocessing used on the original parallel corpus is then applied to the synthetic corpus. We produce backtranslations using **Large** 100K models with the sampling hyperparameters listed in Table 3.

Statistics on generated synthetic data before and after filtering can be found on Table 1.

2.4 Noisy Channel Reranking

We further improve translations by using Noisy Channel Reranking (Yee et al., 2019), which reranks every candidate translation token $\hat{y}_i^{(T)}$ using Bayes' Rule, as follows:

$$P(\hat{y}_i^{(T)}|x;\hat{y}^{(T-1)}) = \frac{P(x|\hat{y}^{(T-1)})P(\hat{y}^{(T-1)})}{P(x)} \quad (2)$$

where $P(\hat{y}_i^{(T)})$ refers to the probability of the ith candidate token at timestep T given source sentence x and current translated tokens $\hat{y}^{(T-1)}$.

All probabilities are parameterized as standard encode-decoder Transformer neural networks: the **Direct Model** $f_{\phi_D}(x,\hat{y}^{(T-1)})$ models $P(\hat{y}_i^{(T)}|x;\hat{y}^{(T-1)})$ or translation between source to target language; the **Channel Model** $f_{\phi_C}(x|\hat{y}^{(T-1)})$ models $P(x|\hat{y}^{(T-1)})$, or the probability of the target translating back into the predicted translation, and; the **Language Model** $f_{\phi_L}(\hat{y}^{(T-1)})$ models $P(\hat{y}^{(T-1)})$ or the probability of the translated sentence to exist. $P(x)$ is generally not modeled since it is constant for all y. This allows us to leverage a strong language model to guide the outputs of the direct model, while using a channel model to constrain the preferred outputs of the language model (which may be unrelated to the source sentence).

Decoding Hyperparameters	
Beam	5
Length Penalty	1.0
k2	5
CM Top-k	500
δ_{ch} en→he	0.2297
δ_{lm} en→he	0.2056
δ_{ch} he→en	0.2998
δ_{lm} he→en	0.2594

Table 4: Hyperparameters used for the final submission models. The values listed for δ_{ch} and δ_{lm} are the ones used for the final submission models. For testing with **Large 100K** models, we set both δ_{ch} and δ_{lm} to 0.3. "k2" refers to the number of candidates sampled per beam while "CM Top-k" refers to the number of most frequent tokens in the channel model's vocabulary that is used as its output vocabulary during decoding to save space.

During beam search decoding, we rescore the top candidates using the following linear combination of all three models:

$$P(\hat{y}_i^{(T)}|x;\hat{y}^{(T-1)})' = \frac{1}{t}log(P(x|\hat{y}^{(T-1)}) + \frac{1}{s}[\delta_{ch}log(P(x|\hat{y}^{(T-1)}) + \delta_{lm}log(P(\hat{y}^{(T-1)}))] \quad (3)$$

where s and t are source / target debiasing terms, δ_{ch} refers to the weight of the channel model, and δ_{lm} refers to the weight of the language model.

For Noisy Channel Reranking, our direct and channel models use the same size and setup at all times (i.e. if the direct model is a **Large** model trained for 100K steps, then the channel model is also a **Large** model trained for 100K steps in the opposite translation direction).

For the language model, we train one **Base**-sized decoder-only Transformer language model for English and one for Hebrew. We concatenate the cleaned data from the parallel corpus with the provided monolingual data for each language to train the LM. We use the same training setup as with translation models, except we use a weight decay of 0.01 and a learning rate of 5e-4.

Hyperparameters used for decoding with Noisy Channel Reranking can be found in Table 4.

2.5 Evaluation

We evaluate our models using two metrics: BLEU (Papineni et al., 2002) and ChrF++ (Popović, 2015), both scored via SacreBLEU[4] (Post, 2018). We develop our models using both the FLORES 200 (Costa-jussà et al., 2022) and NTREX 128 (Federmann et al., 2022) datasets, using the validation sets during training and reporting scores on the test sets.

To benchmark our models' performance, we mainly compare BLEU and ChrF++ against two (unconstrained) models: **`mBART 50 M2M`** (Tang et al., 2020), a 610M-parameter finetuned version of **`mBART`** for many-to-many translation, and **`NLLB 200 MoE`** (Costa-jussà et al., 2022), the full 54.5B-parameter mixture-of-experts version of **`NLLB 200`** for many-to-many translation.

2.6 Hyperparameter Search

To find the best values for δ_{ch} and δ_{lm}, as well as to understand how these parameters affect performance, we use Bayesian Hyperparameter Search. We use the **`Large`** 1M + BT models and run 1000 iterations of search, keeping the length penalty static at 1.0, and sampling both δ_{ch} and δ_{lm} from a gaussian with minimum of 0.01 and maximum of 0.99.

We perform this for both en→he and he→en translation directions and use the results for the final submission model.

3 Results

A summary of our results on benchmarks can be found on Table 5.

3.1 Benchmarking Results

Our submission systems (**`Large` 1M + BT + NC**) exhibit strong performance on both translation directions. On FLORES-200, we achieve 44.24 BLEU for en→he and 42.42 BLEU for he→en. The same systems score 33.77 BLEU for en→he and 36.89 BLEU for he→en on NTREX-128.

We note that these systems perform strongly when compared against much larger, unconstrained baseline models. On FLORES-200, we significantly outperform **`mBART 50 M2M`** on en→he by +24.75 BLEU and on he→en by +11.92 BLEU despite having 67% less parameters (200M vs 610M). Notably, our system performs only slightly worse compared to **`NLLB 200 MoE`** despite having **96% less parameters** compared to the mixture-of-experts model. On FLORES-200, we perform -2.56 BLEU worse on en→he and -6.58 BLEU worse on he→en compared to **`NLLB 200 MoE`**.

3.2 Hyperparameter Search Results

In order to find optimal hyperparameters for both δ_{ch} and δ_{lm}, we ran bayesian hyperparameter search for both at the same time while keeping length penalty static. We plot the results of the hyperparameter search over 1000 iterations in Figure 1.

We observe that performance is optimal when both hyperparameters are set to 0.2∼0.3, making performance increasingly worse as both hyperparameters approach closer to 1. We hypothesize that this signifies the model capturing the original distribution close enough that it does not need much correction or aid from the accompanying language model. Noisy channel reranking, however, is still empirically shown to be useful in this case as guidance from the language model produces better candidates in cases where the direct model may be searching a too-constrained space.

3.3 Ablations

We explored multiple configurations of our submission systems in terms of model size, presence of synthetic data during training, and the use of reranking methods during online decoding. Our results show that each step improves performance directly:

- The initial **`Base` 100K** performs at 39.88 BLEU for en→he on FLORES-200.
- Increasing the size to 200M parameters (**`Large` 100K**) improves performance by +1.38 BLEU.
- Adding backtranslated data (**`Large` 100K + BT**) is by far the most beneficial, improving performance by +2.06 BLEU.
- We then experiment with longer training times (1M iterations for **`Large` 1M + BT**) to adapt to the new dataset size, increasing the score by +0.44 BLEU.
- Finally, using noisy channel reranking (**`Large` 1M + BT + NC**) improves the score by +0.48 BLEU.

[4]SacreBLEU outputs the following signature for evaluation: nrefs:1|case:mixed|eff:no|tok:spm-flores|smooth:exp|version:2.2.1

	FLORES-200				NTREX-128			
	EN → HE		HE → EN		EN → HE		HE → EN	
Model	**BLEU**	**ChrF++**	**BLEU**	**ChrF++**	**BLEU**	**ChrF++**	**BLEU**	**ChrF++**
Base 100K	39.88	56.34	12.06	29.46	31.47	48.32	29.85	52.53
Base 100K + NC	40.22	56.55	38.75	60.52	32.10	48.93	31.86	54.57
Base 100K + BT	41.50	57.46	38.73	60.80	31.27	47.90	34.09	56.10
Base 100K + BT + NC	41.66	57.59	40.43	62.17	32.05	48.62	35.76	57.65
Large 100K	41.26	57.46	39.07	60.06	32.49	48.95	31.08	53.19
Large 100K + NC	41.46	57.64	40.53	61.49	32.80	49.34	33.12	55.16
Large 100K + BT	43.32	58.62	40.91	61.58	32.90	49.11	35.48	56.04
Large 100K + BT + NC	43.26	58.72	41.92	62.64	33.18	49.42	36.79	57.37
Large 1M + BT	43.76	58.29	41.00	61.16	33.35	49.22	35.83	56.02
Large 1M + BT + NC	**44.24**	**59.36**	**42.42**	**62.21**	**33.77**	**49.69**	**36.89**	**56.92**
mBART50 M2M (610M)	19.49	46.7	30.50	55.00	14.80	42.30	27.02	51.21
NLLB 200 MoE (54.5B)	46.80	59.80	49.00	67.40	-	-	-	-

Table 5: Compiled results for all experiments. "BT" refers to the model being trained with backtranslated data in addition to original filtered data. "NC" refers to the use of Noisy Channel Reranking. Evaluation scores for **NLLB 200 MoE** are taken from its official published scores for FLORES-200. We fail to report independent NTREX-128 scores for **NLLB 200 MoE** due to a lack of computational resources.

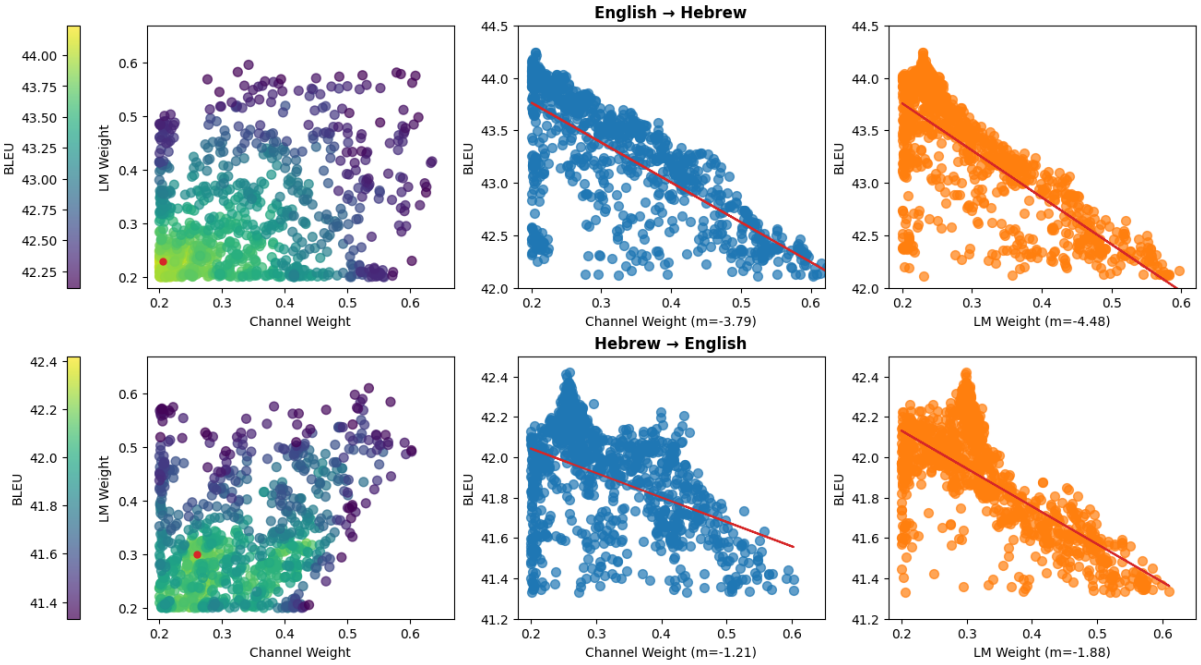


Figure 1: Bayesian hyperparameter search results for δ_{ch} and δ_{lm} while keeping constant length penalty. The leftmost column shows BLEU score against both δ_{ch} and δ_{lm} with the best performing model (**Large** 1M + BT + NC) plotted in red. The middle and rightmost columns show δ_{ch} and δ_{lm} against BLEU, respectively, with their respective regression lines (in red) and regression coefficient (m) in the caption.

Overall, all of our methods improve performance by a total of 4.36 BLEU for the en→he direction on FLORES-200.

We note an interesting jump in performance from **Base 100K** to **Large 1M + BT + NC** on the FLORES-200 he→en direction at +30.36 BLEU. **Base 100K** underperforms at 12.06 BLEU, and we hypothesize that this is due to the model not having enough capacity to embed information from Hebrew, which causes it to greatly benefit from the guidance of a language model during noisy channel reranking.

4 Conclusion

In this paper, we describe our submissions to the WMT 2023 General Translation Task. We participate in two constrained tracks: en→he and he→en.

We submit two monodirectional models based on the Transformer architecture. Both models are trained using a mix of original and synthetic back-translated data, filtered and curated using a comprehensive data processing pipeline that combines embedding-based, heuristic-based, and ratio-based filters. Additionally, we employ noisy channel reranking to improve translation candidates using a language model and a channel model trained in the opposite direction.

On two benchmark datasets, our systems outperform **mBART50 M2M** and perform slightly worse than **NLLB 200 MoE**, both unconstrained systems with significantly more parameters.

Our results show that established best practices still perform strongly on constrained systems without the need for extraneous data sources as is with unconstrained systems for the same translation directions.

Limitations

We benchmark on datasets that are publicly available with permissive licenses for research.

We note that we are unable to study scale properly for translation models due to a lack of stronger compute resources. The same constraint also prevents us from training multiple iterations of the same model with differing random seeds. Our systems' true performance may thus be higher or lower depending on the machine random state at the start of training time.

Lastly, our models are trained on Hebrew, which is a language that we do not speak. We are therefore unable to manually evaluate if the output translations are correct, natural, or semantically sound.

Ethical Considerations

Our paper replicates best practices in data preprocessing, model training, and online decoding for translation models. Within our study, we aim to create experiments that replicate prior work under comparable experimental conditions to ensure fairness in benchmarking.

Given that we do not speak the target language in the paper, we report performance in comparison to other existing models. We do not claim that "strong" performance in a computational setting correlates with good translations from a human perspective.

Lastly, while we do not use human annotators for this paper, the conference (WMT) itself does for human evaluations on the General Translation Task. We disclose this fact and note that annotations (and therefore scores) may be different across many speakers of Hebrew.

References

Dan Bareket and Reut Tsarfaty. 2021. Neural Modeling for Named Entities and Morphology (NEMO2). *Transactions of the Association for Computational Linguistics*, 9:909–928.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Jan Christian Blaise Cruz and Charibeth Cheng. 2021. Improving large-scale language models and resources for filipino. *arXiv preprint arXiv:2111.06053*.

Jan Christian Blaise Cruz and Lintang Sutawika. 2022. Samsung research philippines-datasaur ai's submission for the wmt22 large scale multilingual translation task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1034–1038.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015a. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015b. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Lintang Sutawika and Jan Christian Blaise Cruz. 2021. Data processing matters: Srph-konvergen ai's machine translation system for wmt'21. *arXiv preprint arXiv:2111.10513*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jie Yang and Yue Zhang. 2018. NCRF++: An open-source neural sequence labeling toolkit. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Kyra Yee, Nathan Ng, Yann N Dauphin, and Michael Auli. 2019. Simple and effective noisy channel modeling for neural machine translation. *arXiv preprint arXiv:1908.05731*.

NAIST-NICT WMT'23 General MT Task Submission

Hiroyuki Deguchi[1,2] **Kenji Imamura**[2] **Yuto Nishida**[1]
Yusuke Sakai[1] **Justin Vasselli**[1] **Taro Watanabe**[1]
[1]Nara Institute of Science and Technology
[2]National Institute of Information and Communications Technology
{deguchi.hiroyuki.db0, nishida.yuto.nu8,sakai.yusuke.sr9,
vasselli.justin_ray.vk4, taro}@is.naist.jp
kenji.imamura@nict.go.jp

Abstract

In this paper, we describe our NAIST-NICT submission to the WMT'23 English ↔ Japanese general machine translation task. Our system generates diverse translation candidates and reranks them using a two-stage reranking system to find the best translation. First, we generated 50 candidates each from 18 translation methods using a variety of techniques to increase the diversity of the translation candidates. We trained seven models per language direction using various combinations of hyperparameters. From these models we used various decoding algorithms, ensembling the models, and using kNN-MT (Khandelwal et al., 2021). We processed the 900 translation candidates through a two-stage reranking system to find the most promising candidate. In the first step, we compared 50 candidates from each translation method using DrNMT (Lee et al., 2021) and returned the candidate with the best score. We ranked the final 18 candidates using COMET-MBR (Fernandes et al., 2022) and returned the best score as the system output. We found that generating diverse translation candidates improved translation quality using the well-designed reranker model.

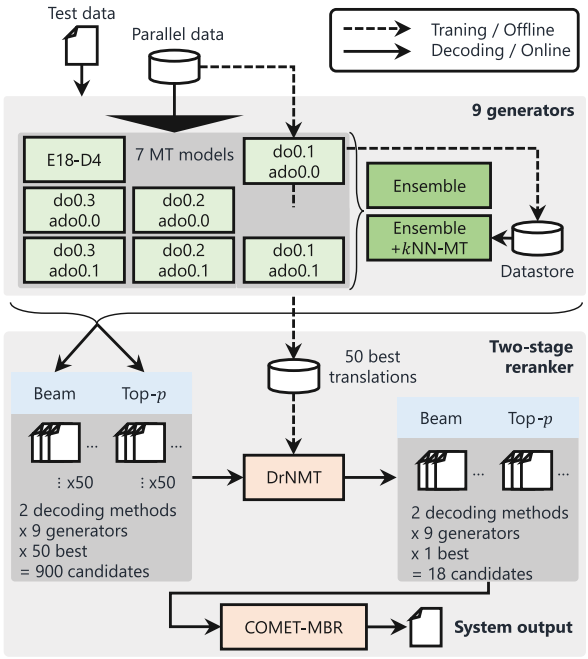


Figure 1: Overview of our system. "E18-D4" denotes "18-layer encoder and 4-layer decoder", and "do" and "ado" denote "dropout" and "dropout after applying attention softmax", respectively.

1 Introduction

We participated in the WMT'23 general machine translation task for English-to-Japanese (En-Ja) and Japanese-to-English (Ja-En) translation. Our team aimed to improve translation performance using only the provided parallel data. Our system generates diverse translation candidates and reranks them using a two-stage reranking system to find the best translation.

Figure 1 shows an overview of our system. We trained 7 Transformer (Vaswani et al., 2017) NMT models per language direction using various combinations of hyperparameters. The translation generator consists of 9 instances: 7 MT models, the ensemble model, and a kNN-MT (Khandelwal et al., 2021) system that interpolates tokens from retrieved examples using the ensemble model. The generator generates the 50-best translations each from two decoding methods: beam search and top-p sampling. This combination allows the generator to find diverse translation candidates. Next, the 900 candidates (9 generators $\times$ 2 decoding methods $\times$ 50 best) are passed to our two-stage reranker to find the best translation. The first step of reranking uses DrNMT (Lee et al., 2021) to rerank the 50-best translation candidates to select the 1-best translation from each of the 18 generator and decoding method combinations. DrNMT is trained to maximize the BLEU (Papineni et al., 2002) score, whereas we use the second step reranking to find the highest COMET (Rei et al., 2020) score expectation from the remaining candidates. The 18

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 110–118
December 6–7, 2023.

candidates from the first step are reranked using COMET-MBR (Fernandes et al., 2022) to select the best translation that is returned by the system.

Our experiments show that our two-stage reranker outperforms the BLEU, chrF, and COMET scores by DrNMT alone, and the BLEU and chrF scores by COMET-MBR alone in both En-Ja and Ja-En translation tasks on wmttest2022 (Kocmi et al., 2022).

2 Preprocessing

For the training data, we used the provided bilingual parallel data, which included JParaCrawl v3 (Morishita et al., 2020), News Commentary v18.1, Wiki Titles v3, WikiMatrix, the Japanese-English Subtitle Corpus (Pryzant et al., 2017), the Kyoto Free Translation Task Corpus (Neubig, 2011), and the Web Inventory of Transcribed and Translated Talks (Cettolo et al., 2012). We did not backtranslate the monolingual data due to resource constraints for training MT models and a reranker model.

As the English translation of the Japanese-English Subtitle Corpus was only available in lowercase, we trained a Moses truecaser (Koehn et al., 2007) using the other corpora to add capitalization to the subtitle corpus. After truecasing, the first letter of each sentence was capitalized using detruecasing to produce sentence-case English text that matched the casing in the other corpora.

We cleaned the data by removing duplicate lines and applying language filtering. Because much of the training data were crawled from the internet, we used fasttext (Joulin et al., 2016a,b) to predict the language of each sentence and removed sentences that were not predicted to be in the correct language. This helped to reduce noise in the dataset by removing sentences with garbage tokens.

We tokenized text into subword units using sentencepiece (Kudo and Richardson, 2018). Since our system generates many candidates using multiple models, we preliminary measured the generation speed and selected the number of vocabulary with the fastest decoding. Our initial experiments demonstrated that when the target language was Japanese, a vocabulary size of 32k resulted in fewer tokens needing to be generated, which increased the translation speed. However, when the target language was English, a vocabulary size of 16k was faster than an English vocabulary of 32k. Therefore, we trained separate dictionaries for English and Japanese, with the English-side dictionary containing nearly 16k tokens and the Japanese-side containing nearly 32k tokens. The character coverage of the tokenizers also varied between languages. We trained the English tokenizer with 100% character coverage, whereas character coverage for Japanese was 99.98%.

After subword segmentation, we removed all sentences shorter than one token or longer than 250 tokens. We also removed all sentences in which the number of tokens in one language was more than double the number of tokens in the translation, i.e., the ratio of tokens between the source and target was >2.0. The number of sentence pairs before/after preprocessing is shown in Table 1.

	#sentence pairs
No filter	33,875,242
+ deduplicate	29,940,444
++ language filter	29,279,161
+++ length filter	27,880,378

Table 1: Number of sentence pairs in the training data after each preprocessing step.

Generator: MT model	
Architecture	Transformer big
Embedding dimension	1,024
FFN inner dimension	8,192
Dropout (do)	0.1
Attention dropout (ado)	0.0
Loss function	label smoothed cross entropy
Label smoothing	$\epsilon = 0.1$
Optimizer	Adam ($\beta_1 = 0.9, \beta_2 = 0.98$)
Learning rate (LR)	1e-3
LR scheduler	inverse square root
Warm-up steps	4,000
Global batch size	Roughly 512,000 tokens
Training steps	60,000
Reranker: DrNMT	
Architecture	XLM-R large
Classifier dropout	0.2
Loss function	(Section 3.2.1)
Optimizer	Adam ($\beta_1 = 0.9, \beta_2 = 0.98$)
Learning rate (LR)	5e-5
LR scheduler	polynomial decay
Warm-up steps	8,000
Global batch size	512 sentences * 50 hypotheses

Table 2: Hyperparameters of the models we trained.

3 Translation System

3.1 Generator

The generator generates diverse translation candidates from multiple models and multiple decod-

ing methods. The generator consists of seven MT models, an ensemble of the seven models, and the ensemble enhanced with kNN-MT (Khandelwal et al., 2021) for a total of 9 instances.

3.1.1 MT models

The 7 MT models are trained from the provided parallel data. Our MT model with the default setting is shown in Table 2. Six of the seven models vary from the default setting only in dropout and attention dropout, while the last varies the number of layers. Our model has two types of dropouts whose values are varied: "dropout (do)" and "attention dropout (ado)". The dropout (do) is applied to the token embedding layer and the outputs of the sub-layers within each layer, i.e., the outputs of the attention layers and feed-forward network. The attention dropout (ado) is applied after softmax to the attention weights, i.e., before multiplying the values. Six models are trained with varying dropouts, one for each combination of $\mathrm{do} = \{0.1, 0.2, 0.3\}$ and $\mathrm{ado} = \{0.0, 0.1\}$. In addition to the models that vary dropout, we trained a deep-shallow model (Kasai et al., 2021), which has 18 encoder layers and 4 decoder layers. For each model, we averaged the parameters of the last 10 checkpoints (10,000 training steps).

3.1.2 kNN-MT

Datastore construction kNN-MT (Khandelwal et al., 2021) requires a datastore to be constructed to store the translation examples to be accessed during decoding. Let $\boldsymbol{x} = (x_1, \ldots, x_{|\boldsymbol{x}|}) \in \mathcal{V}_X^{|\boldsymbol{x}|}$ and $\boldsymbol{y} = (y_1, \ldots, y_{|\boldsymbol{y}|}) \in \mathcal{V}_Y^{|\boldsymbol{y}|}$ denote a source sentence and target sentence, respectively, where $|\cdot|$ is the length of the sequence, and $\mathcal{V}_X$ and $\mathcal{V}_Y$ are the vocabularies of the source language and target language, respectively. The datastore for kNN-MT consists of translation examples in the form of key–value pairs, as shown in Figure 2. Each target token y_t from the translation examples is stored in the datastore with a d-dimensional key ($\in \mathbb{R}^d$), which is the representation of the translation context $(\boldsymbol{x}, \boldsymbol{y_{<t}})$ obtained from the decoder of the pre-trained NMT model. The datastore $\mathcal{M} \subseteq \mathbb{R}^d \times \mathcal{V}_Y$ is formally defined as a set of tuples as follows:

$$\mathcal{M} = \{(f(\boldsymbol{x}, \boldsymbol{y}_{<t}), y_t) \mid (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}, 1 \leq t \leq |\boldsymbol{y}|\}, \quad (1)$$

where $\mathcal{D}$ denotes parallel data and $f : \mathcal{V}_X^{|\boldsymbol{x}|} \times \mathcal{V}_Y^{t-1} \to \mathbb{R}^d$ returns the intermediate representation of the final decoder layer from the source sentence and prefix target tokens. The representation used as the key vector is the vector that is passed into the final feed-forward layer (Khandelwal et al., 2021).

In our system, we used the model trained with the default settings (as seen in Table 2) to obtain the keys for the datastore.

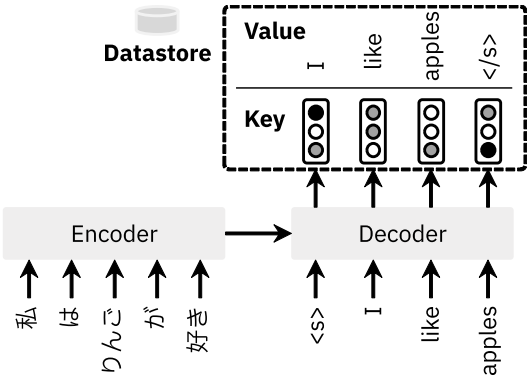


Figure 2: Datastore construction.

kNN index	
Implementation	FAISS
Index	`IndexIVFPQ`
# of entries	
Ja (En-Ja)	732,222,393
En (Ja-En)	836,254,078
# of centroids	131,072
# of bits in PQ	8 bits
# of sub-vectors in PQ	$M = 64$
Vector pre-transform	OPQ (Ge et al., 2014)
Decoding	
# of retrieved tokens	$k = 64$
Temperature of $p_{k\mathrm{NN}}$	$\tau = 100$
Weight for $p_{k\mathrm{NN}}$	$\lambda = 0.1$
# of probed clusters	32

Table 3: Hyperparameters of our kNN indexes and kNN-MT.

kNN index To search the k-nearest-neighbor tokens efficiently, we used FAISS (Johnson et al., 2019). For the kNN indexes, we used `faiss.IndexIVFPQ` which consists of an inverted file index (IVF) that performs k-means clustering and product quantization (PQ) (Jégou et al., 2011) which divides a vector into M sub-vectors and performs vector quantization in each subspace. Note that in IVFPQ, the codewords of PQ are learned from the residual vectors from the centroids of the IVF. Additionaly, we used optimized PQ (OPQ) (Ge et al., 2014) to reduce the quantiza-

tion error of PQ. The hyperparameters of our kNN indexes are summarized in Table 3.

Decoding During decoding, kNN-MT retrieves the k-nearest-neighbor key–value pairs $\{(\boldsymbol{k}_i, v_i)\}_{i=1}^{k} \subseteq \mathbb{R}^d \times \mathcal{V}_Y$ from the datastore $\mathcal{M}$ using the query vector $f(\boldsymbol{x}, \boldsymbol{y}_{<t})$ at timestep t. Next, $p_{k\mathrm{NN}}$ is calculated as follows:

$$p_{k\mathrm{NN}}(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t}) \propto \sum_{i=1}^{k} \mathbb{1}_{y_t=v_i} \exp \frac{-\|\boldsymbol{k}_i - f(\boldsymbol{x}, \boldsymbol{y}_{<t})\|_2^2}{\tau}, \quad (2)$$

where τ is the temperature parameter for $p_{k\mathrm{NN}}$. Then, kNN-MT generates the output probability by computing the linear interpolation between the kNN and MT probabilities, $p_{k\mathrm{NN}}$ and p_{MT}, respectively:

$$P(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t}) = \lambda p_{k\mathrm{NN}}(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t}) + (1-\lambda) p_{\mathrm{MT}}(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t}). \quad (3)$$

kNN-MT with the ensemble model kNN-MT is typically used with a single model, whereas in our system, we obtain the output probability for each token by interpolating between the kNN probability and the probability from the ensemble model. The output probability from the ensemble kNN-MT is formulated by defining p_{MT} in Equation 3 as follows:

$$p_{\mathrm{MT}}(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t}; \boldsymbol{\theta}) = \frac{1}{|\boldsymbol{\theta}|}(p_{\mathrm{MT}}(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t}; \theta_1) + \ldots + p_{\mathrm{MT}}(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t}; \theta_{|\boldsymbol{\theta}|}), \quad (4)$$

where $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_{|\boldsymbol{\theta}|}\}$ denotes the parameters of the trained MT models; $|\boldsymbol{\theta}| = 7$ in our system. The kNN-MT decoding interpolated between the token distribution of the retrieved translation context tokens and the full ensemble of models. As such, the weight assigned to the kNN token distribution was kept small so as not to overpower the information from the ensemble. We used $\lambda = 0.1$ and $\tau = 100$ in the kNN-MT decoding shown in Table 3.

3.1.3 Decoding algorithms

From each model, we output the 50 best hypotheses generated using beam search with a beam width of 50. For diversity, we generated another 50 hypotheses using top-p sampling with $p = 0.7$ and a beam width of 50. We formed an ensemble of models to produce two more sets of 50 hypothesis sentences from beam search and top-p sampling.

3.2 Reranker

We use a two-stage reranker consisting of an intra-system reranker, which selects the best of the 50 hypotheses from each system, and an inter-system reranker, which selects the best hypothesis from the 18 remaining candidate translations.

3.2.1 DrNMT

Discriminative reranking for NMT (DrNMT) (Lee et al., 2021) is a discriminative model that learns to predict the distributions of the evaluation scores of a set of translation hypotheses given a source sentence. DrNMT is similar to a quality estimation model (Zerva et al., 2022), but it is optimized to distinguish the better translation from hypotheses generated from a single system. In addition, it cannot be used for comparing inter-systems because the weights for features are tuned using the translation hypotheses of the development set. We used BLEU (Papineni et al., 2002) as the evaluation metric for this first-stage reranker.

Model The DrNMT model takes as input a source sentence $\boldsymbol{x} \in \mathcal{V}_X^{|\boldsymbol{x}|}$ concatenated with a hypothesis translation $\boldsymbol{y}^{(j)} \in \mathcal{V}_Y^{|\boldsymbol{y}^{(j)}|}$. The DrNMT model passes this into XLM-R (Conneau et al., 2020), which is a multilingual pre-trained encoder. The hidden state of the `[CLS]` token then represents the combination of the source and hypothesis and is converted into a scalar score by the classification head of RoBERTa (Liu et al., 2019). We used an input dimension of 1,024, a hidden dimension of 768, and output dimension of 1. The activation function for the classification head is tanh.

Objective The objective function minimizes the KL-divergence between the DrNMT model distribution and the distribution of BLEU scores of the n-best hypotheses; that is, the objective function $\mathcal{L}(\theta)$ is as follows:

$$\mathcal{L}(\theta) = \mathrm{KL}[p_T \parallel p_M] = -\sum_{j=1}^{n} p_T\left(\boldsymbol{y}^{(j)}, \boldsymbol{y}^*\right) \log p_M\left(\boldsymbol{y}^{(j)}|\boldsymbol{x}; \theta\right), \quad (5)$$

where n denotes the number of translation hypotheses, and p_M and p_T denote the distributions of the DrNMT model and BLEU scores, respectively. $\boldsymbol{y}^*$ denotes the reference translation of $\boldsymbol{x}$. The BLEU scores are normalized using min-max scaling and the distribution of the BLEU scores is emphasized

using the temperature coefficient T. In this paper, we use $T = 0.5$.

Training We trained the DrNMT model using the 50 best translation hypotheses generated by the model with the default configuration for each source sentence over the entire training set, i.e., 28M source sentences. The model is trained using early stopping, which selects the checkpoint with the maximum BLEU score in the validation set.

Tuning The score of the first-stage reranker is a weighted sum of the DrNMT model score, translation model score, and length penalty. This combination of scores is similar to minimum error rate training (Och, 2003). The weights that maximize the BLEU score of the validation set were learned and used.

Implementation We used the implementation published in FAIRSEQ[1]. Note that this implementation uses SACREBLEU (Post, 2018) to compute the BLEU scores. We modified the published code of DrNMT to change the SACREBLEU tokenizers according to the target language because the published implementation always calls the English tokenizer.

3.2.2 COMET-MBR

COMET-MBR (Fernandes et al., 2022) performs minimum Bayes risk (MBR) decoding (Kumar and Byrne, 2004; Eikema and Aziz, 2020; Müller and Sennrich, 2021; Eikema and Aziz, 2022) using a COMET (Rei et al., 2020, 2022) model trained on direct assessments. A translation $\hat{\boldsymbol{y}}^{\text{MAP}} \in \mathcal{V}_Y^{|\boldsymbol{y}^*|}$ is typically generated using maximum-a-posteriori (MAP) decoding as follows:

$$\hat{\boldsymbol{y}}^{\text{MAP}} = \underset{\boldsymbol{y} \in \mathcal{Y}}{\operatorname{argmax}} \log p(\boldsymbol{y}|\boldsymbol{x}), \tag{6}$$

where $\mathcal{Y} \subseteq \bigcup_{i=1}^{\infty} \mathcal{V}_Y^i$ is the search space of target sentences. In MBR decoding, instead of finding the most probable translation, the goal is to find the translation that minimizes the Bayes risk as follows:

$$\hat{\boldsymbol{y}}^{\text{MBR}} = \underset{\boldsymbol{h} \in \bar{\mathcal{Y}}}{\operatorname{argmax}} \underbrace{\mathbb{E}_{\boldsymbol{y}' \sim p(\boldsymbol{y}|\boldsymbol{x})}[u(\boldsymbol{y}', \boldsymbol{h})]}, \tag{7}$$
$$\approx \frac{1}{m} \sum_{j=1}^{m} u(\boldsymbol{y}^{(j)}, \boldsymbol{h})$$

where $\bar{\mathcal{Y}} \subseteq \mathcal{Y}$ is a set of translation hypotheses and $u : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is the utility function. In this paper, we used COMET[2] (Rei et al., 2020, 2022) as utility function u. Note that we share the hypotheses $\bar{\mathcal{Y}}$ and the sample set for expectation estimation $\{\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(m)}\}$, except for $\boldsymbol{h}$, i.e., $\{\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(m)}\} = \bar{\mathcal{Y}} \setminus \{\boldsymbol{h}\}$. Thus, given a candidate set, the computational complexity of MBR decoding is in the order of $\mathcal{O}(m^2)$, which results in a slower inference speed when m is large.

3.2.3 Two-stage reranking

We applied two-stage reranking with DrNMT and COMET-MBR, which allowed us to use each model for the task it was trained to handle best, to optimize for two metrics and to reduce the inference speed of reranking.

In the first stage, DrNMT (Lee et al., 2021) is used to prune the 50 candidates for each candidate set generated from each of the 18 combinations of decoding methods and generators. As DrNMT is trained to rerank the n-best candidates from a single model, it is ideally suited to the task of reranking the candidates generated with the same combination of model and decoding method, i.e., within a system. In the second stage, COMET-MBR (Fernandes et al., 2022) is used to select the system output from the 18 candidate translations selected by DrNMT.

We use COMET-MBR to rerank the best outputs of each system because COMET was trained on translation scores from the output of various models from previous WMT translation tasks, making it well suited to inter-system comparisons. Each of the two stages is trained to optimize a different metric: Stage one uses BLEU, which evaluates surface forms, whereas stage two uses COMET, which evaluates semantics. Additionally, the inference speed of COMET-MBR makes it time-consuming for large candidate sets, but pruning with DrNMT, which performs inference in a single forward computation, reduces the computational cost.

4 Experimental Results

We evaluated the translation performance of our system on wmttest2022 (Kocmi et al., 2022). We measured the BLEU and chrF scores using SACREBLEU, and the COMET score using `Unbabel/wmt22-comet-da`. The models of our

[1] https://github.com/facebookresearch/fairseq/tree/main/examples/discriminative_reranking_nmt

[2] https://huggingface.co/Unbabel/wmt22-comet-da

Method	# of cands.	En-Ja			Ja-En		
		BLEU	chrF	COMET	BLEU	chrF	COMET
1-best of the ensemble	1	25.5	34.0	86.4	23.1	48.0	80.9
DrNMT	50	26.7	34.7	86.6	23.7	48.4	81.1
COMET-MBR	900	26.1	35.4	**90.5**	22.0	48.0	**84.1**
DrNMT+COMET-MBR (ours)	900	**27.1**	**35.6**	88.4	**24.4**	**49.3**	82.4
DrNMT+Oracle-COMET-DA	900	30.5	39.1	90.2	29.0	53.7	85.5

Table 4: Experimental results of our system on wmttest2022. "# of cands." denotes the number of candidates generated by the translation generator. The **bold scores** indicate the best scores in each translation direction.

generator were trained using FAIRSEQ (Ott et al., 2019). We used KNN-SEQ[3] (Deguchi et al., 2023) for kNN-MT generation built on top of FAIRSEQ. The first stage of our reranker, DrNMT, was also built using FAIRSEQ, whereas COMET-MBR was built using COMET (Rei et al., 2020).

Table 4 shows the results of our system. In the table, the translation candidates of "1-best of the ensemble" were generated using the ensemble model without kNN-MT using beam search decoding. The candidates of "DrNMT" were generated using the ensemble model and the 50-best translations were obtained using beam search decoding. As DrNMT uses the log probability of an MT model for inference, it cannot compare candidates generated by different MT models or generation methods. The results show that DrNMT not only improved the BLEU scores but also the chrF and COMET scores from the 1-best translation, despite only being trained to maximize the BLEU score. "COMET-MBR" reranks all candidates, i.e., 900 translations (= 9 generators $\times$ 2 decoding methods $\times$ 50 best candidates). COMET-MBR achieved the highest COMET scores for both En-Ja and Ja-En, but the BLEU and chrF scores were not improved for Ja-En, and the inference speed of COMET-MBR with 900 translation candidates was slow. Our primary system used "DrNMT+COMET-MBR" described in Section 3.2.3. This method obtained higher scores for all metrics compared with using DrNMT alone in both translation directions, in addition to the highest BLEU and chrF scores overall. To summarize, our results show that using the rerankers appropriately as intra- and inter-system rerankers is effective for improving translation quality. DrNMT+Oracle-COMET-DA is the oracle performance of the second stage reranker, i.e., the score computed by the largest COMET-DA score for candidates after reranking the 50-best of each system using DrNMT (first stage reranker). Our DrNMT+COMET-MBR scores underperformed the oracle performance, and we leave its improvement for future work.

In addition, we investigated which hypothesis was selected as the system output in DrNMT+COMET-MBR. Figure 3 shows the percentages of counts selected as the system output. In the figure, when the system output comes from multiple hypotheses, i.e., duplicated hypotheses are selected, each hypothesis is counted as selected. The results show that the hypotheses generated by beam search of the ensemble and ensemble+kNN-MT models were selected as the system outputs roughly 40% in En-Ja and 50% in Ja-En. Thus, half of the system outputs were not selected from hypotheses generated from the ensemble model using beam search. Therefore, it can be said that "DrNMT+COMET-MBR" outperformed "DrNMT" by selecting from the hypotheses generated by various generators and various decoding methods.

5 Conclusion

In this paper, we described our submission as a joint team of NAIST and NICT (NAIST-NICT) to the WMT'23 general MT task. We participated in this task in the En-Ja and Ja-En translation directions. We built our system using a diverse translation generator and two-stage reranker. In future work, we will investigate qualitatively how translation diversity contributes to translation quality.

Limitations

A limitation of our system is its reliance on large computation resources. As our system generates 50 candidates using two decoding methods from

[3] https://github.com/naist-nlp/knn-seq

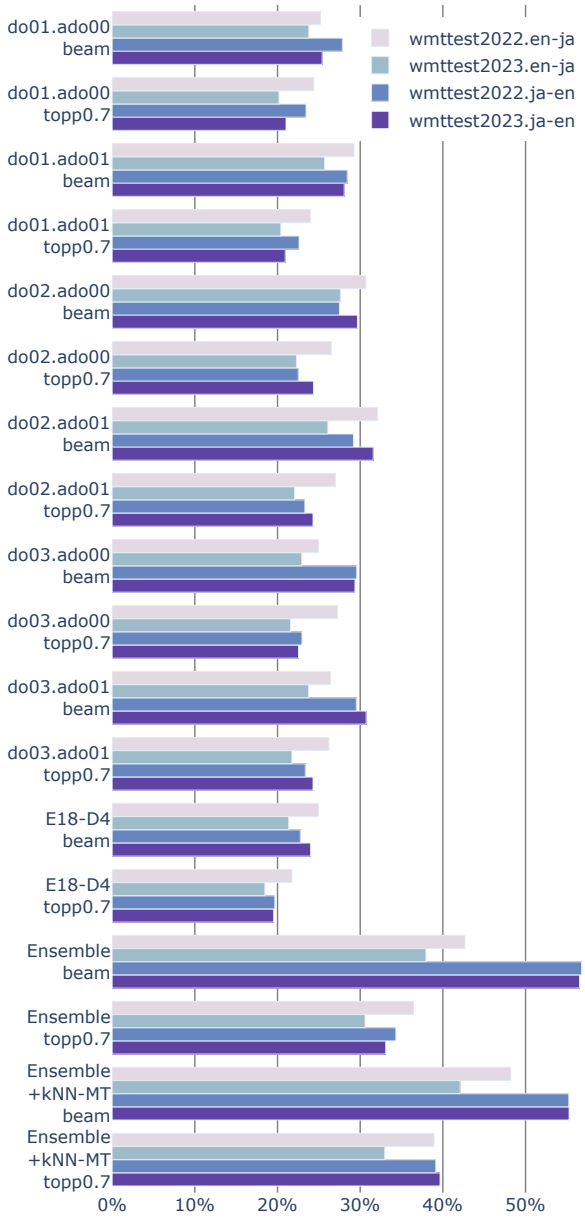


Figure 3: Percentages of counts selected as the system output by COMET-MBR.

each of the nine generators, it requires significant resources. The beam size of 50 is larger than most machine translators and requires more computing power (memory and time).

Note that the reranking approach cannot output translations of higher quality than those translated by the generators.

Ethics Statement

Our system did not restrict the training data and the translator's outputs. Therefore, similar to other translation systems, it may generate factually inaccurate translations.

References

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy. European Association for Machine Translation.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Hiroyuki Deguchi, Hayate Hirano, Tomoki Hoshino, Yuto Nishida, Justin Vasselli, and Taro Watanabe. 2023. knn-seq: Efficient, extensible knn-mt framework. *arXiv preprint arXiv:2310.12352*.

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Bryan Eikema and Wilker Aziz. 2022. Sampling-based approximations to minimum Bayes risk decoding for neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10978–10993, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

Tiezheng Ge, Kaiming He, Qifa Ke, and Jian Sun. 2014. Optimized product quantization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(4):744–755.

Hervé Jégou, Matthijs Douze, and Cordelia Schmid. 2011. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Jungo Kasai, Nikolaos Pappas, Hao Peng, James Cross, and Noah Smith. 2021. Deep encoder, shallow decoder: Reevaluating non-autoregressive machine translation. In *International Conference on Learning Representations*.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest neighbor machine translation. In *International Conference on Learning Representations (ICLR)*.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Ann Lee, Michael Auli, and Marc'Aurelio Ranzato. 2021. Discriminative reranking for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7250–7264, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2020. JParaCrawl: A large scale web-based English-Japanese parallel corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3603–3609, Marseille, France. European Language Resources Association.

Mathias Müller and Rico Sennrich. 2021. Understanding the properties of minimum Bayes risk decoding in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 259–272, Online. Association for Computational Linguistics.

Graham Neubig. 2011. The Kyoto free translation task. http://www.phontron.com/kftt.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Reid Pryzant, Yongjoo Chung, Dan Jurafsky, and Denny Britz. 2017. JESC: japanese-english subtitle corpus. *CoRR*, abs/1710.10639.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

CUNI at WMT23 General Translation Task: MT and a Genetic Algorithm

Josef Jon and **Martin Popel** and **Ondřej Bojar**
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{jon,popel,bojar}@ufal.mff.cuni.cz

Abstract

This paper presents the contributions of Charles University teams to the WMT23 General translation task (English to Czech and Czech to Ukrainian translation directions). Our main submission, CUNI-GA, is a result of applying a novel n-best list reranking and modification method on translation candidates produced by the two other submitted systems, CUNI-Transformer and CUNI-DocTransformer (document-level translation only used for the $en \rightarrow cs$ direction). Our method uses a genetic algorithm and MBR decoding to search for optimal translation under a given metric (in our case, a weighted combination of ChrF, BLEU, COMET22-DA, and COMET22-QE-DA). Our submissions are first in the constrained track and show competitive performance against top-tier unconstrained systems across various automatic metrics.

1 Introduction

Our submission for this year's WMT General translation task (Kocmi et al., 2023) is based on the previous submissions of our team (Popel et al., 2019, 2022) and MBR decoding in combination with genetic algorithm (GA). We describe the method in separate work (Jon and Bojar, 2023). The main goal of our submission is to find out whether our approach improves the translation quality perceived by humans. For this reason, we submitted both the base system translations and the mutated and reranked (i.e. GA-processed) translations for the human evaluation.

As all the parts of the approach are described in detail in the mentioned papers (as well as all the related work), we will restrict ourselves to providing a short overview of the main points in Section 2. In Section 3, we describe the datasets, tools and parameters used to obtain results presented in Section 4. Finally, we draw conclusions from the results.

2 Methods

Our submissions make use of two features that are not typical for current MT systems: document-level context and translation refinement through a genetic algorithm.

2.1 Document level translation

We use document-level NMT for the $en \rightarrow cs$ direction. The approach is described in Popel et al. (2019). Since all the training data for this direction have document boundaries, a document-level training set is created by extracting all sequences of consecutive sentences with at most 3000 characters. The final training set consists of pairs of such examples, where both sides have the same number of sentences. Sentences are separated by a special token. We also use Block backtranslation (Popel, 2018; Popel et al., 2020; Gebauer et al., 2021; Jon et al., 2022a).

2.2 Genetic algorithm

Our approach (Jon and Bojar, 2023) utilizes MBR decoding (Goel and Byrne, 2000; Kumar and Byrne, 2004; Amrhein and Sennrich, 2022; Freitag et al., 2021; Müller and Sennrich, 2021; Jon et al., 2022b) in conjunction with the genetic algorithm (GA) (Fraser, 1957; Bremermann, 1958; Holland, 1975). By merging and mutating translations generated by an MT system, we aim to find the best translation under a specific metric. This is a new strategy for creating translation candidates in NMT. We illustrate one iteration of the whole process in Figure 1. The top, yellow part shows the steps that are the same as in simple reranking. We have an initial population of candidates, for example, n-best list produced by an MT model, that is scored by *fitness function*, in our case, a sum of MBR decoding scores using an MT evaluation metric and QE scores. At this point, for reranking, the process would stop after selecting the best-scoring translation candidate. In GA, we continue by splitting

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 119–127
December 6–7, 2023.

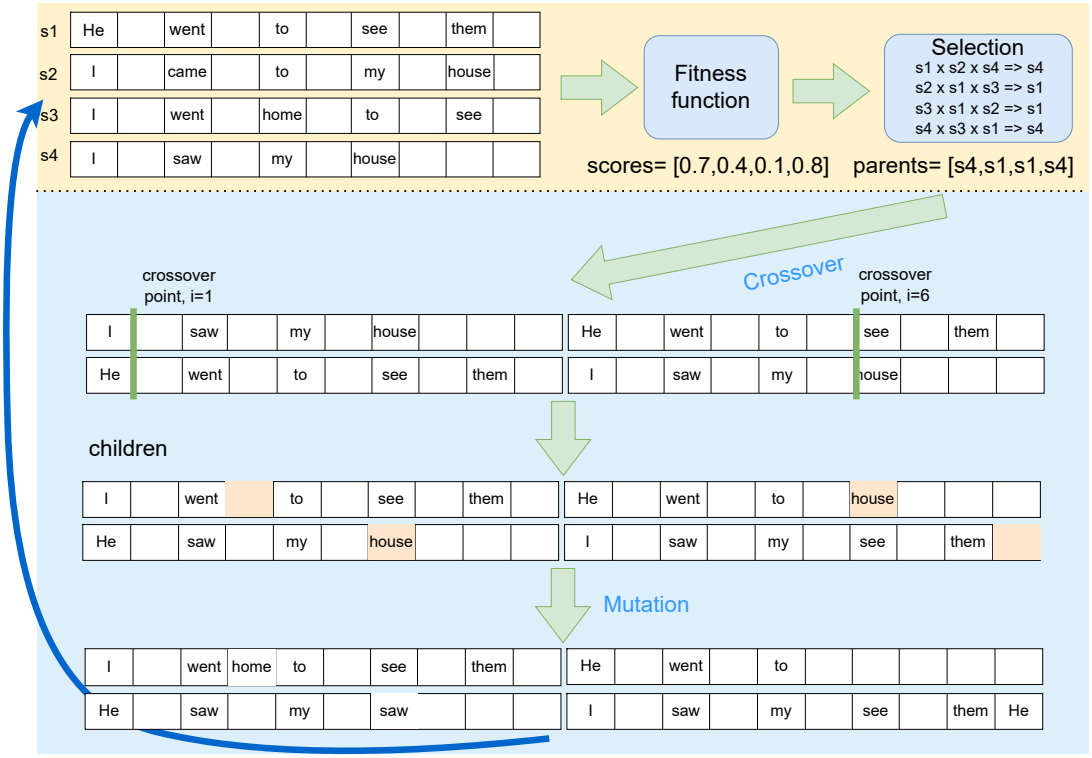


Figure 1: One iteration of the GA algorithm for a population of 4 individuals. The steps with a yellow background are equivalent to simple reranking, the steps with blue background introduce the operations of the genetic algorithm. Figure taken from Jon and Bojar (2023)

a well-scoring subset of the candidate sentences at random points and reattaching them in a different order, by a process called cross-over. These combined candidates are mutated at random places, meaning some of the tokens are either deleted or replaced by different tokens from a set of suitable candidate tokens. Also, new tokens can be added this way. These modifications result in a new population of translation candidates and the whole process is repeated from the start. A more detailed description of our approach is available in Jon and Bojar (2023).

MBR decoding NMT models generate a probability distribution over potential translations for a specified input sentence. The widely used method to derive the ultimate translation from this distribution is "maximum-a-posteriori" (MAP) decoding. However, the computational demands of precise MAP decoding lead to the adoption of approximations like beam search, referenced by Koehn et al. (2003). Recent literature, such as Stahlberg and Byrne (2019) and Meister et al. (2020), has shed light on several constraints of MAP and proposed alternatives.

MBR decoding is one such alternative. It uses a utility function to select the translation, aiming to minimize expected loss or *risk*. Typically, MT metrics are employed as these utility functions. In practice, candidate translations produced by the MT model are used as an approximation of the set of all possible translations. In such case, if we only use purely reference-based metrics (like BLEU), MBR decoding becomes a consensus decoding, where the chosen candidate is the one closest to all the others. However, novel MT metrics also take source sentence into account, so the process is more complex than a simple search for the most average translation. The MBR decoding has seen renewed interest with the introduction of the new generation of metrics (Amrhein and Sennrich, 2022; Freitag et al., 2021; Müller and Sennrich, 2021; Jon et al., 2022b).

3 System description

Our models are based on submissions of our team from previous years (Popel et al., 2022, 2019). We resubmit those (*CUNI-Transformer* and *CUNI-DocTransformer* submissions) and we also submit an additional translation: the outputs of these models combined, mutated and rescored by the GA

described in Section 2.2 (*CUNI-GA* submission).

3.1 Tools and data

All our submissions are constrained, using only the training data provided by the task organizers, specifically the CzEng 2.0 (Kocmi et al., 2020) corpus. We used English to Czech `newstest-18` and `newstest-22` as validation sets for the genetic algorithm approach. Due to the computational requirements of our method, we only evaluate the first 150 sentences of each test set. We didn't run any validation experiments for GA in the $cs \rightarrow uk$ language pair, we used the same parameters as for $en \rightarrow cs$. We have only translated the general translation test set using GA, the test suits translations for *CUNI-GA* are copied from the *CUNI-DocTransformer* submission.

3.2 Models

We use Transformer models. For the dev set experiments, we use same models as Jon and Bojar (2023) (i.e. `transformer-big` using MarianNMT (Junczys-Dowmunt et al., 2018) with default hyperparameters). For the final submissions, the models are the same as in last year's submissions: Popel et al. (2022) for $cs \rightarrow uk$ and Popel et al. (2019) for $en \rightarrow cs$.

3.3 GA parameters

We refrained from searching for the optimal values of GA parameters due to the significant computational demands of our method.

For the results on the validation set, we used exactly the settings described by Jon and Bojar (2023), i.e. Transformer model trained on the CzEng 2.0 (Kocmi et al., 2020) corpus in $cs \rightarrow en$ direction (i.e. the opposite direction to the task). We used beam search with size 20 to produce a 20-best list and sampled an additional 20 translations from the model to create an initial population of 40 candidates, which we copied 50 times to obtain a population size of 2000.

We used different NMT models (see Section 3.2) and a different number of initial sentences for the shared task submissions. For the $cs \rightarrow uk$ direction, the starting population consists of the top 35 hypotheses produced by beam search from the two models described in Popel et al. (2022) (top-10 from the *CUNI-Transformer-inca-roman* and top-25 from the *CUNI-Transformer* model).[1] This set is replicated 50 times, leading to a total population of 1750 candidates. For $en \rightarrow cs$ we use a concatenation of n-best lists with beam sizes 4 and 10 from both *CUNI-DocTransformer* and *CUNI-Transformer* (28 candidates in total), also copied 50 times over, resulting in population size of 1400. To combine document-level and sentence-level translations, we re-split the translated documents back into sentences.

To choose parents for the succeeding generation, we use tournament selection with $n = 3$. These parents are then merged at a crossover rate of $c = 0.1$. The mutation rate, for altering non-empty genes (i.e. tokens) to other non-empty genes m, is $1/l$, where l denotes the chromosome's (chromosome is a sequence of tokens, representation of one translation candidate) length[2] For transitions from an empty to a non-empty gene (i.e. addition of a word) and vice versa (i.e. deletion), the rate is $\frac{m}{10}$. The GA runs for 250 and 130 generations for $cs \rightarrow uk$ and $en \rightarrow cs$, respectively.

3.4 Metrics

The translations are evaluated by the following metrics: ChrF (Popović, 2015), BLEU (Papineni et al., 2002), BLEURT (Sellam et al., 2020), multiple versions of COMET (Rei et al., 2020, 2021, 2022b,a,c) and UniTE (Wan et al., 2022). We abbreviate some of the longer metrics' names further in the text in order to save space.[3]

For both BLEU and ChrF, we utilize SacreBLEU (Post, 2018). In all experiments, ChrF uses a $\beta = 2$ setting (ChrF2). We rely on the original implementations for COMET,[4] BLEURT,[5] and UniTE[6] scores.

4 Results

This section presents automatic metric scores on validation sets and the official test set.

4.1 English to Czech

The first translation direction is English to Czech, where we submitted the outputs of our older sentence-level (*CUNI-Transformer*) and document-level (*CUNI-DocTransformer*) systems, as well as

[1] The *CUNI-Transformer-inca-roman* uses preprocessing using romanization and inline casing (Popel et al., 2022).

[2] See Jon and Bojar (2023) for a more detailed description.

[3] CMT20 (wmt20-comet-da), CMT21 (wmt21-comet-mqm), CMTH22 (eamt22-cometinho-da), QE20 (wmt20-comet-qe-da-v2), QE22 (wmt22-cometkiwi-da), BLEURT (BLEURT-20), UniTE (UniTE-MUP)

[4] `https://github.com/Unbabel/COMET`

[5] `https://github.com/google-research/bleurt`

[6] `https://github.com/NLP2CT/UniTE`

Method	Fitness	ChrF	BLEU	CMT20	CMT21	CMTH22	QE20	BLEURT	UniTE	% new
baseline	-	56.7	30.1	0.5007	0.0399	0.5017	0.2477	0.7078	0.3018	0
Reranking	CMT20	57.4	31.2	0.5853	0.0409	0.5390	0.2930	0.7193	0.3413	0
Reranking	CMT20+QE20+BLEU	**57.5**	**31.2**	0.5983	0.0417	**0.5596**	0.3620	**0.7255**	0.3686	0
GA	CMT20	56.2	28.4	0.6247	0.0410	0.5382	0.2893	0.7177	0.3366	52
GA	CMT20+QE20+BLEU	**57.5**	29.5	**0.6266**	**0.0429**	0.5403	**0.4198**	0.7174	**0.3946**	70

Table 1: Comparison of the scores of baseline MT output, reranked output, and GA-modified output. The last column shows the percentage of finally selected best translations that were not present in the initial population (i.e. they were newly created by the GA operations). Table from Jon and Bojar (2023).

Model	w_{CMT}	w_{QE}	w_{BLEU}	w_{chrF}	chrF	BLEU	CMT20	CMT21	CMTH22	QE20	CMT22	BLEURT	UniTE	New
Baseline	-	-	-	-	56.6	30.1	0.500	0.040	0.504	0.244		0.707	0.301	0.00
	0.15	0.15	0.35	0.35	57.2	29.8	0.619	**0.043**	0.542	0.401	0.856	0.715	0.384	0.64
	0.1	0.1	0.4	0.4	**57.4**	30.0	0.616	**0.043**	0.541	0.403	0.856	0.714	0.385	0.63
	0.25	0.25	0.25	0.25	**57.4**	29.8	0.616	**0.043**	0.541	0.410	0.857	0.713	0.388	0.64
	0.2	0.2	0.3	0.3	57.3	29.6	0.619	**0.043**	0.540	0.406	0.857	0.715	0.388	0.64
CMT20	0.4	0.2	0.2	0.2	57.2	**30.7**	0.629	**0.043**	**0.549**	0.388	0.856	**0.720**	0.384	0.51
	0.4	0.3	0.1	0.2	**57.4**	30.2	0.630	**0.043**	0.548	0.405	0.857	0.718	0.384	0.65
	0.4	0.4	0.1	0.1	57.1	29.0	0.631	**0.043**	0.542	0.427	**0.859**	0.716	**0.389**	0.68
	0.5	0.5	0	0	55.2	25.1	**0.633**	**0.043**	0.514	**0.470**	0.856	0.705	0.372	0.86
	1	0	0	0	56.8	29.7	0.614	0.041	0.533	0.289	0.844	0.712	0.336	0.51
	0.15	0.15	0.35	0.35	57.5	32.0	0.601	0.042	0.560	0.332	0.858	0.729	0.388	0.27
	0.1	0.1	0.4	0.4	**57.7**	**32.2**	0.602	0.042	**0.562**	0.331	0.858	0.730	0.392	0.28
	0.25	0.25	0.25	0.25	57.5	32.0	0.601	0.042	0.560	0.330	0.857	0.729	0.388	0.29
CMT22	0.2	0.2	0.3	0.3	57.5	32.0	0.601	0.042	0.561	0.331	0.858	0.730	0.394	0.32
	0.4	0.2	0.2	0.2	57.2	31.5	0.593	0.042	0.550	0.326	0.857	0.727	0.370	0.25
	0.4	0.3	0.1	0.2	**57.7**	32.1	0.597	0.042	0.555	0.332	0.857	0.728	0.386	0.27
	0.4	0.4	0.1	0.1	57.6	32.0	0.606	0.042	0.560	0.334	0.859	0.730	0.393	0.29
	0.5	0.5	0	0	**57.7**	31.7	**0.620**	**0.043**	**0.562**	**0.359**	**0.866**	**0.731**	**0.406**	0.57
	1	0	0	0	56.8	29.8	0.570	0.042	0.528	0.328	0.863	0.714	0.344	0.49

Table 2: Scores of translations on the first 150 sentences of `newstest-18` created by GA. The fitness metric is a weighted sum of COMET, COMET-QE, BLEU and chrF, with weight shown in columns 2 to 5. The first column shows which version of COMET and COMET-QE was used. Higher is better for all the metrics. The best results for each metric are bold.

Model	w_{CMT}	w_{QE}	w_{BLEU}	w_{chrF}	chrF	BLEU	CMT20	CMT21	CMTH22	QE20	CMT22	BLEURT	UniTE	New
Baseline	-	-	-		68.3	44.9	0.738	0.045	0.751	0.357	0.876	0.785	0.540	0.00
	0.15	0.15	0.35	0.35	68.4	43.0	0.777	0.047	0.777	0.464	0.890	0.787	0.607	0.52
	0.1	0.1	0.4	0.4	68.6	43.5	0.779	0.047	0.779	0.464	0.891	0.787	0.609	0.51
	0.25	0.25	0.25	0.25	68.3	43.0	0.785	0.047	**0.783**	0.469	0.892	0.789	**0.617**	0.52
	0.2	0.2	0.3	0.3	68.5	43.3	0.780	0.047	0.777	0.465	0.891	0.787	0.610	0.52
CMT20	0.4	0.2	0.2	0.2	**68.6**	**44.2**	0.778	0.047	0.773	0.441	0.887	**0.791**	0.586	0.33
	0.4	0.3	0.1	0.2	68.2	43.0	0.785	0.047	0.777	0.470	0.891	0.789	0.612	0.49
	0.4	0.4	0.1	0.1	67.7	42.1	**0.787**	**0.047**	0.777	**0.485**	**0.892**	0.788	0.614	0.55
	0.5	0.5	0	0	65.1	36.4	0.782	0.047	0.747	0.514	0.887	0.771	0.574	0.77
	1	0	0	0	67.9	42.1	0.772	0.046	0.760	0.386	0.880	0.785	0.552	0.36
	0.15	0.15	0.35	0.35	68.8	45.1	0.771	0.047	0.794	0.417	0.890	0.799	0.604	0.25
	0.1	0.1	0.4	0.4	68.8	45.1	0.772	0.047	**0.795**	0.417	0.890	0.798	0.605	0.25
	0.25	0.25	0.25	0.25	68.8	44.8	0.774	0.047	0.792	0.418	0.890	0.799	0.603	0.27
CMT22	0.2	0.2	0.3	0.3	68.9	45.3	0.772	0.047	0.794	0.417	0.890	0.799	0.604	0.25
	0.4	0.2	0.2	0.2	68.9	45.1	0.771	0.047	0.794	0.408	0.889	0.795	0.602	0.22
	0.4	0.3	0.1	0.2	**69.1**	**45.7**	0.772	0.047	0.794	0.410	0.889	0.798	0.607	0.25
	0.4	0.4	0.1	0.1	68.8	45.2	0.771	0.047	0.792	0.420	0.890	0.798	0.603	0.25
	0.5	0.5	0	0	68.6	43.6	**0.782**	**0.047**	0.793	**0.431**	**0.893**	**0.800**	**0.612**	0.46
	1	0	0	0	68.2	43.5	0.762	0.046	0.778	0.401	0.890	0.788	0.576	0.39

Table 3: Scores of translations on the first 150 sentences of `newstest-22` created by GA. The fitness metric is a weighted sum of COMET, COMET-QE, BLEU and chrF, with weight shown in columns 2 to 5. The first column shows which version of COMET and COMET-QE was used. Higher is better for all the metrics. The best results for each COMET version are bold.

a combination and modification of both using our GA approach.

GA vs. reranking Jon and Bojar (2023) provide a comparison of the genetic algorithm approach to a simple reranking using the same objective metrics. In that work, a sum of CMT20, QE20 and BLEU is used as the fitness metric. The results are copied in Table 1. The baseline translations are obtained via beam search. The same work also shows that for UniTE, CMT22, CMT21-MQM held-out metrics[7], GA significantly outperforms simple reranking with the same objective metric. However, BLEURT, CMTH22 and chrF seem to favor reranking only.

For our current work, we ran additional experiments. We use a weighted sum of COMET, COMET-QE, chrF and BLEU as the objective (fitness) metric. We compare older and newer versions of both COMET and COMET-QE, represented by CMT20/QE20 and CMT22/QE22, respectively. Since the objective metrics lose their relevance for evaluation once we optimize for them, a set of held-out metrics is selected to better estimate the translation quality. The results for the first 150 sentences of `newstest18` are presented in Table 2, and the scores for the first 150 sentences of `newstest22` are presented in Table 3.

We vary the weights of the different fitness metrics to see the effect on the held-out metrics (columns $\mathbf{w_{CMT}}$, $\mathbf{w_{QE}}$, $\mathbf{w_{BLEU}}$ and $\mathbf{w_{chrF}}$). The last column shows a portion of cases where the final selected candidate was not part of the initial population, the other columns show values of the respective scores.

We see an interesting difference between CMT22/QE22 and CMT20/QE20. While optimizing only for CMT20 or CMT20+QE20 hurts other scores greatly (for example UniTe and BLEURT), optimizing solely for CMT22+QE22 does not have such an adverse effect on other metrics. We hypothesize multiple factors play a role in this. One of them might be the better robustness of the newer versions, which are designed to deal better with hallucinations and unexpected target tokens that could be introduced by the GA. CMT20 and especially QE20 were previously shown to be partially insensitive to this kind of errors (Guerreiro et al., 2023), but they could be detected by the other metrics, hence the lower scores.

Final submission Overall, the results suggest the best choice is to simply average CMT22 and QE22 scores ($w_{CMT} = 0.5$, $w_{QE} = 0.5$). We did not have the complete evaluation at hand by the time of the submission, so we used weights $w_{CMT} = 0.4$, $w_{QE} = 0.4$, $w_{BLEU} = 0.1$ and $w_{chrF} = 0.1$ for the submitted test set translation. We use a completely different NMT system than in the dev set experiments to create the initial population for the submission, as described in 3.3.

We show the automatic scores of all the submissions on the test set in Table 4. The *CUNI-GA* submission outperforms both the base submissions *CUNI-Transformer* and *CUNI-DocTransformer* across all metrics. It ranks comparably to the best unconstrained system using COMET, but lags behind in chrF and BLEU.

We analyzed the percentages of the final submitted translated sentences that were present in some of the initial n-best lists and the percentage of novel sentences, created by the GA. We show these results in Table 5. We see that 21.7% of the final submitted sentences are new, not contained in any of the initial n-best lists, but rather created by the GA mutation and crossover operations.

4.2 Czech to Ukrainian

We also ran the GA on a concatenation of n-best lists produced by the two $cs \rightarrow uk$ models, see Popel et al. (2022) for details on these systems. We used beam size 10 for the *CUNI-Transformer-inca-roman* model and beam size 25 for the *CUNI-Transformer* model, resulting in 35 initial candidate sentences. We did not perform any parameter tuning on the validation set, we used the same parameters as for the $en \rightarrow cs$ submission. We present the automatic metrics results on the test set in Table 6. Our submissions outperform the only other constrained system and are competitive with the unconstrained systems, scoring best in COMET and 2nd in chrF and BLEU. For COMET and chrF, GA outperforms the unmodified baseline translation, while in BLEU, the baseline scores slightly better.

Again, we show what is the percentage of final best translations selected for submission contained in either of the initial n-best lists and the percentage of new translations, created by GA operations, in Table 7. 35.1% of the final submitted translations are novel.

[7]Means metrics not used as a part of the fitness function. Note that these metrics are not completely independent, they can be still linked to the fitness metrics by spurious correlations caused by data and model architecture similarity

System	COMET
ONLINE-W	91.8
CUNI-GA	90.8
ONLINE-B	89.9
GPT4-5shot	89.4
ONLINE-A	88.4
CUNI-DocTransformer	88.3
GTCOM_Peter	87.7
ONLINE-M	87.4
Lan-BridgeMT	87.3
CUNI-Transformer	87.2
NLLB_Greedy	87.1
ONLINE-Y	87.0
NLLB_MBR_BLEU	86.9
ONLINE-G	85.9
ZengHuiMT	85.4

System	chrF
ONLINE-W	76.3
ONLINE-B	70.4
ZengHuiMT	67.5
ONLINE-A	66.3
CUNI-GA	65.9
GTCOM_Peter	65.4
CUNI-DocTransformer	65.1
ONLINE-Y	64.6
CUNI-Transformer	63.9
Lan-BridgeMT	63.8
ONLINE-G	63.7
ONLINE-M	63.2
GPT4-5shot	62.3
NLLB_Greedy	60.0
NLLB_MBR_BLEU	59.1

System	BLEU
ONLINE-W	59.4
ONLINE-B	50.1
ONLINE-A	43.4
CUNI-GA	43.3
ZengHuiMT	43.1
CUNI-DocTransformer	42.5
GTCOM_Peter	42.3
CUNI-Transformer	41.4
ONLINE-Y	40.8
Lan-BridgeMT	40.7
ONLINE-G	39.6
ONLINE-M	39.6
GPT4-5shot	37.8
NLLB_Greedy	35.9
NLLB_MBR_BLEU	35.1

Table 4: Results of automatic evaluation on $en \rightarrow cs$ testset. Unconstrained systems are indicated with a grey background. Coincidentally, all three $en \rightarrow cs$ unconstrained systems are our submissions described in this paper. CUNI-GA is better than the two baselines according to all three metrics.

	doc-4	doc-10	sent-4	sent-10	new
contains	36.3%	42.1%	31.4%	52.8%	21.7%
unique	2%	6.8%	0.7%	17.5%	
merge	50.0%		53.5%		
u-merge	24.7%		33.3%		

Table 5: Percentages of final best scoring in CUNI-GA English to Czech submission sentences by the initial n-best list they are contained in (doc-4 denotes document-level, beam size 4 and so on). The first row shows how many sentences from the final translation were present in the respective n-best list, while the last column shows the percentage of completely new sentences, that were not present in any of the lists. The second row looks at the percentages of final sentences that are uniquely in exactly one of the lists. The last two rows show the same for merged doc-level and sent-level lists, i.e. we concatenated both beam sizes for each into one list.

5 Future work

Our setting allows many straightforward modifications to potentially improve the results of our method. First of all, MBR decoding works well on a large, diverse set of initial candidates, obtained for example by sampling. In our experiments, we only use short n-best lists produced by beam search. An additional benefit stemming from the diversity of the initial candidates is a more dive diverse set of possible tokens for replacement mutations.

Second, we did not run any search for the parameters of the GA process (crossover and mutation rates, number of generations, population size, selection method), due to the large computational costs of this approach. We believe a set of better parameters could be found easily by, for example, a grid search. Finally, the metrics used for the fitness function are combined by a simple weighted sum. Multi-criterion genetic algorithms can be explored for a better approach to combine multiple evaluation scores for the translations.

Also, reranking and modifying the translations on a sentence level can introduce inconsistencies previously mitigated by using document-level MT, losing the advantages of document-level processing. Deutsch et al. (2023) show that using sentence-level metrics for whole document-level segments might be a viable option for avoiding this issue.

6 Conclusion

We confirm that using MBR decoding in combination with a genetic algorithm can improve scores in selected evaluation metrics, while creating original novel translations. We show that our systems are competitive in both submitted language pairs, winning among constrained systems based on automated evaluation metrics.

7 Acknowledgements

This work was partially supported by GAČR EXPRO grants NEUREM3 (19-26934X) and LUSyD (20-16819X), TAČR grant EdUKate (TQ01000458), by the Grant Agency of Charles University in Prague (GAUK 244523) and by the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID:90254). It has been using data and tools provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

System	COMET
CUNI-GA	90.9
GPT4-5shot	90.8
ONLINE-W	89.4
GTCOM_Peter	88.9
ONLINE-B	88.8
ONLINE-A	88.2
CUNI-Transformer	88.0
ONLINE-G	87.7
MUNI-NLP	87.0
ONLINE-Y	86.5
NLLB_Greedy	86.3
NLLB_MBR_BLEU	86.3
Lan-BridgeMT	86.0

System	chrF
GPT4-5shot	61.0
CUNI-GA	57.9
GTCOM_Peter	57.6
CUNI-Transformer	57.4
MUNI-NLP	57.0
Lan-BridgeMT	55.7
ONLINE-W	55.0
ONLINE-B	54.7
ONLINE-A	54.4
ONLINE-G	53.7
ONLINE-Y	53.4
NLLB_Greedy	52.5
NLLB_MBR_BLEU	52.3

System	BLEU
GPT4-5shot	32.8
CUNI-Transformer	30.2
GTCOM_Peter	29.8
CUNI-GA	29.5
MUNI-NLP	28.3
Lan-BridgeMT	27.5
ONLINE-W	26.8
ONLINE-B	25.7
ONLINE-A	25.4
NLLB_MBR_BLEU	25.1
NLLB_Greedy	24.9
ONLINE-G	24.8
ONLINE-Y	24.2

Table 6: Results of automatic evaluation on $cs \rightarrow uk$ testset. Unconstrained systems are indicated with a grey background. CUNI-GA is better than CUNI-Transformer according to COMET and chrF, but worse according to BLEU.

	CT-inca-roman-10	CT-25	new
contains	17%	58.5%	35.1%
unique	6.4%	48%	

Table 7: Percentages of final best scoring sentences by the initial n-best list they are contained in, the meaning of the rows is the same as in Table 5. CT=CUNI-Transformer.

References

Chantal Amrhein and Rico Sennrich. 2022. Identifying weaknesses in machine translation metrics through minimum bayes risk decoding: A case study for comet.

Hans J Bremermann. 1958. *The evolution of intelligence: The nervous system as a model of its environment*. University of Washington, Department of Mathematics.

Daniel Deutsch, Juraj Juraska, Mara Finkelstein, and Markus Freitag. 2023. Training and meta-evaluating machine translation evaluation metrics at the paragraph level.

Alex S Fraser. 1957. Simulation of genetic systems by automatic digital computers ii. effects of linkage on rates of advance under selection. *Australian Journal of Biological Sciences*, 10(4):492–500.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2021. Minimum bayes risk decoding with neural metrics of translation quality.

Petr Gebauer, Ondřej Bojar, Vojtěch Švandelík, and Martin Popel. 2021. CUNI systems in WMT21: Revisiting backtranslation techniques for English-Czech NMT. In *Proceedings of the Sixth Conference on Machine Translation*, pages 123–129, Online. Association for Computational Linguistics.

Vaibhava Goel and William J Byrne. 2000. Minimum bayes-risk automatic speech recognition. *Computer Speech & Language*, 14(2):115–135.

Nuno M. Guerreiro, Elena Voita, and André Martins. 2023. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075, Dubrovnik, Croatia. Association for Computational Linguistics.

John H. Holland. 1975. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, MI. Second edition, 1992.

Josef Jon and Ondřej Bojar. 2023. Breeding machine translations: Evolutionary approach to survive and thrive in the world of automated evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2191–2212, Toronto, Canada. Association for Computational Linguistics.

Josef Jon, Martin Popel, and Ondřej Bojar. 2022a. CUNI-bergamot submission at WMT22 general translation task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 280–289, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Josef Jon, Martin Popel, and Ondřej Bojar. 2022b. CUNI-Bergamot Submission at WMT22 General Translation Task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 280–289, Abu Dhabi. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme

Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Martin Popel, and Ondrej Bojar. 2020. Announcing czeng 2.0 parallel corpus with over 2 gigawords. *arXiv preprint arXiv:2007.03006*.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Clara Meister, Ryan Cotterell, and Tim Vieira. 2020. If beam search is the answer, what was the question? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2173–2185, Online. Association for Computational Linguistics.

Mathias Müller and Rico Sennrich. 2021. Understanding the properties of minimum Bayes risk decoding in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 259–272, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Martin Popel. 2018. CUNI transformer neural MT system for WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 482–487, Belgium, Brussels. Association for Computational Linguistics.

Martin Popel, Jindřich Libovický, and Jindřich Helcl. 2022. CUNI systems for the WMT 22 Czech-Ukrainian translation task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 352–357, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Martin Popel, Dominik Macháček, Michal Auersperger, Ondřej Bojar, and Pavel Pecina. 2019. English-Czech systems in WMT19: Document-level transformer. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 342–348, Florence, Italy. Association for Computational Linguistics.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11(4381):1–15.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, José G.C. de Souza, Pedro G. Ramos, André F.T. Martins, Luisa Coheur, and Alon Lavie. 2022b. Searching for COMETINHO: The little metric that could. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 61–70, Ghent, Belgium. European Association for Machine Translation.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T.

Martins. 2022c. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Felix Stahlberg and Bill Byrne. 2019. On NMT search errors and model errors: Cat got your tongue? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3356–3362, Hong Kong, China. Association for Computational Linguistics.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek F. Wong, and Lidia S. Chao. 2022. UniTE: Unified Translation Evaluation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

SKIM at WMT 2023 General Translation Task

Keito Kudo ♢, **Takumi Ito** ♢, **Makoto Morishita** ♠, **Jun Suzuki** ♢
♢Tohoku University ♠NTT Communication Science Laboratories

Abstract

The SKIM team's submission used a standard procedure to build ensemble Transformer models, including base-model training, back-translation of base models for data augmentation, and retraining of several final models using back-translated training data. Each final model had its own architecture and configuration, including up to 10.5B parameters, and substituted self- and cross-sublayers in the decoder with a cross+self-attention sublayer (Peitz et al., 2019). We selected the best candidate from a large candidate pool, namely 70 translations generated from 13 distinct models for each sentence, using an MBR reranking method using COMET and COMET-QE (Fernandes et al., 2022). We also applied data augmentation and selection techniques to the training data of the Transformer models.

1 Introduction

This paper provides a system description of submissions by our team, called SKIM[1], at WMT-2023. We took part in English to Japanese (En→Ja) and Japanese to English (Ja→En) General Machine Translation tracks (Kocmi et al., 2023). We specifically participated in the constrained track, which places restrictions on the available data and pre-trained models.

The trial of this year's submissions is a reranking part. Our submission system consists of multiple translation models, followed by a reranking module (Kobayashi, 2018) based on COMET (Rei et al., 2022a) and COMET-QE (Rei et al., 2021). This reranking approach serves to identify and select high-quality translations from the hypothesis candidate set generated by multiple translation models. Among the Transformer-based translation models, we also incorporated a large Transformer model with 10.5B parameters. We also applied data augmentation techniques based on our previous year's system (Morishita et al., 2022b). We briefly describe the system overview, including the experimental results that could not be submitted.

[1]The team name is an anagram of the first letters of the authors' last names.

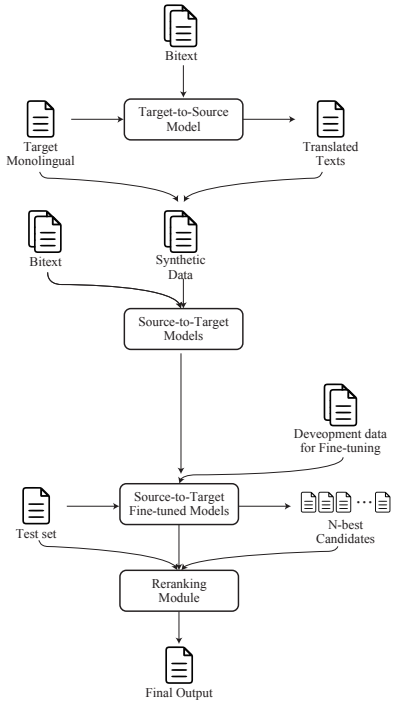


Figure 1: System overview.

2 System Overview

An overview of our submission system is shown in Figure 1. Following the development process used for last year's system (Morishita et al., 2022b), we used Transformer (Vaswani et al., 2017) as the model architecture and conducted pre-training and fine-tuning. In the pre-training phase, we used both a synthetic dataset created by back translation (Sennrich et al., 2016) and the provided bitext dataset. Here, we refer to the target-to-source translation model to generate this synthetic dataset as the initial translation model. Furthermore, we conducted fine-tuning on the translation models derived from pre-training using high-quality bitext

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 128–136
December 6–7, 2023.

Initial Translation Model	
Subword Size	32,000
Architecture	Transformer (big) with FFN size of 4,096
Optimizer	Adam ($\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 1 \times 10^{-8}$)
Learning Rate Schedule	Inverse square root decay
Warmup Steps	4,000
Max Learning Rate	0.001
Dropout	0.3
Gradient Clip	1.0
Batch Size	1,280,000 tokens
Number of Updates	50,000 steps
Averaging	Save a checkpoint every 200 steps and average the last eight
Implementation	`fairseq` (Ott et al., 2019)
Pre-training Configuration	
Subword Size	64,000
Architecture	(See Table 4)
Optimizer	Adam ($\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 1 \times 10^{-8}$)
Learning Rate Schedule	Inverse square root decay
Warmup Steps	4,000
Max Learning Rate	0.001
Dropout	0.3 / 0.1
Gradient Clip	0.1 / 1.0
Batch Size	1,024,000 / 64,000 tokens
Max. Num. of Updates	60,000 / 100,000 (stoped at 64,000)
Averaging	Save a checkpoint every 2,000 steps and average the last ten
Implementation	`fairseq` (Ott et al., 2019)
Fine-tuning Configuration	
Subword Size	Identical to Pre-training Configuration
Architecture	(See Table 4)
Optimizer	Adam ($\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 1 \times 10^{-8}$)
Learning Rate Schedule	Fixed
Warmup Steps	N/A
Max Learning Rate	0.00001
Dropout	0.3 / 0.1
Gradient Clip	1.0
Batch Size	16,000 / 14,400 tokens
Number of Updates	400 / 200
Averaging	Save a checkpoint every ten steps and average the last ten
Implementation	`fairseq` (Ott et al., 2019)

Table 1: List of hyper-parameters. We used the initial translation model for creating synthetic data, pre-training configuration to construct pre-training models described in Section 4.2, and fine-tuning configuration to construct models for submission. Note that we used slightly different settings for 10.5B models in a few parameters. We show their settings at the righthand side of the slash mark (/). We used several different model configurations for ensembling. See Table 4 for more details.

datasets (i.e., development data provided by the organizers). When developing last year's submission system, we found that fine-tuning with clean data enhanced translation quality more effectively than domain adaptation. Therefore, we used a similar fine-tuning approach for this year's submission system. By using these datasets, we trained multiple Transformer-based translation models with heterogeneous configurations. During the inference phase, we translated the source sentences using these translation models individually and selected the final translation results using a subsequent reranking process. As reranking, we tried two methods: one used COMET-QE and the other used COMET-MBR (Fernandes et al., 2022) extended to the outputs of multiple models.

Corpus	w/o Filtering	w/Filtering
JParaCrawl v3.0	25.7 M	25.0 M
WikiMatrix	3.89 M	3.64 M
JESC	2.80 M	2.57 M
Wiki Titles v3	757 K	327 K
KFTT	440 K	371 K
TED Talks	242 K	224 K
NewsCommentary v18	3.8 K	3.7 K

Table 2: Number of sentence pairs in bitext corpus.

3 Dataset Construction

3.1 Provided Data

Bitext Corpus We used all the provided bitext corpora: JParaCrawl v3.0 (Morishita et al., 2022a), News Commentary v18, Wiki Titles v3, WikiMatrix, Japanese-English Subtitle Corpus (JESC) (Pryzant et al., 2018), The Kyoto Free Translation Task (KFTT) Corpus (Neubig, 2011), and TED Talks (Cettolo et al., 2012). We filtered out the potentially noisy pairs using the straightforward parallel corpus filtering methods, just as we did with last year's system (Morishita et al., 2022b). Table 2 shows the size of each dataset with/without filtering. Compared to the previous year, the organizers updated the NewsCommentary, resulting in an increase of 1.8 K sentences.

Monolingual Corpus We also used the following provided monolingual data: News Crawl, News Commentary, and Common Crawl. We back-translated the monolingual sentences using a target-to-source model (i.e., an initial translation model) trained only with the provided bitext dataset, as described in Section 3.2, and used them as synthetic data (Sennrich et al., 2016).

	#sent. pairs	#subwords (JA)	#subwords (EN)
En→Ja	587 M	12.9 B	15.0 B
Ja→En	681 M	17.2 B	16.7 B

Table 3: Statistics of synthetic data used for pre-training.

3.2 Building Pre-Training Data

Synthetic Data Construction To augment the training data, we constructed synthetic data by applying the initial translation model trained with bitext to the monolingual data. As a preprocessing step, we truecased[2] both the bitext and monolingual data. We then tokenized the data into subwords using the `Sentencepiece` tool (Kudo and Richardson, 2018) with the unigram language model option.

We set the vocabulary size to 64,000, the same as the previous year's submission. To integrate insights from the method to create vocabulary for recent large-language models (Touvron et al., 2023), we activated the "byte_fallback" and "split_digits" options. Through preliminary experiments, we confirmed that activating these options leads to enhanced translation performance. As our initial translation model, we used the identical initial translation model we used for last year's submission system (Morishita et al., 2022b). The detailed hyperparameters are described in the initial translation model section of Table 1. Finally, we respectively translated 3.3 B (English) and 1.4B (Japanese) monolingual sentences.

Data Cleaning For both the provided bitext and synthetic data, we carried out cleaning based on a combination of sentence embeddings and handcrafted rules.

For both the bitext and synthetic data, we removed the too-long sentences (>500 characters) and using the `langid`[3] toolkit, removed the sentences that were identified as not being written in English or Japanese.

For the synthetic data, we further applied a sentence embedding-based filtering approach. We took advantage of LaBSE (Feng et al., 2022) to embed the Japanese and English sentences into the same embedding space. We then scored and ranked the parallel sentence pairs based on the cosine similarity of their sentence embeddings. We subsequently filtered out the following items from the synthetic data:

- Duplicated sentence pairs
- Sentences with over 150 words[4] or single words with over 40 characters
- Sentences where the ratio between the word and the character count is > 12
- Sentences that contain invalid Unicode characters
- Sentence pairs where the source/target word ratio exceeds 4
- sentence pairs where the source/target length ratio exceeds 6
- sentence pairs where the source and target sentences are identical
- sentence pairs where the cosine similarity is greater than 0.96[5]

Finally, we respectively selected the top 587M and 681M (approximately) sentences, respectively, from the translated 1.4 B and 3.3 B monolingual sentences as the En→Ja and Ja→En synthetic data for the rank orders. Table 3 shows the statistics of the synthetic data used for our pre-training.

3.3 Fine-Tuning Data

As mentioned in Section 2, during the development of last year's submission system, we found that fine-tuning the model with clean data was more effective for improving translation quality than domain adaptation. Following this finding, we used the WMT'20 test set, WMT'20 development set, WMT'21 test set and WMT'22 test set as clean data for fine-tuning. The WMT'20 test and development sets were all used as clean data. However, for the WMT'21 and WMT'22 test sets, only the opposite language direction data were used (i.e., only Ja→En data were used as clean data for the En→Ja models) because these data were used for development and evaluation. The clean data included 9,002 sentences for En→Ja and 9,026 sentences for Ja→En.

4 Primary Translation Module

We trained several Transformer models for the reranking in the decoding phase. We describe the

[2] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/recaser/truecase.perl

[3] https://github.com/saffsd/langid.py

[4] We tokenized the Japanese sentences using MeCab (Kudo, 2006) with the IPA dictionary. Note that this tokenization is for this cleaning purpose only.

[5] We found that sentence pairs with high cosine similarities can be noisy; for example, the source and target sentences are sometimes identical. Thus, we removed them from the training data.

details of the models in this section. Furthermore, alongside the newly trained models, we reused the primary translation models from the previous year's submission system (Morishita et al., 2022b).

4.1 Model Configuration

We independently trained models with heterogeneous model configurations. Our configuration has several notable characteristics: a cross+self-attention mechanism and a large number of parameters (i.e., 10.5B). In the following sections, we describe the details of the configurations.

Cross+Self-Attention Mechanism We introduced a cross+self-attention mechanism (Peitz et al., 2019) to the Transformer decoder. This mechanism was expected to reduce the model parameters and provide faster training while maintaining the translation performance. In this approach, we eliminated the decoder's cross-attention layer and unified the self-attention and cross-attention into a single attention layer. Specifically, the self-attention layer within the Transformer decoder simultaneously performs the cross-attention calculation by concatenating the output from the encoder's final layer to the query and key matrices.

Suppose Q, K, and V are the query, key, and value matrices, respectively; H_{enc} is the matrix form of concatenating all the output vectors of the encoder's final layer; W_q, W_k, W_v are the weight matrices for the query, key, and value, respectively; and d_k denotes the dimension of the key matrix. It is then formulated as follows:

$$\begin{aligned} &\text{Attention}(Q, K, V, H_{enc}) = \\ &\qquad \text{softmax}\left(\frac{Q_{concat} K_{concat}^T}{\sqrt{d_k}}\right) V' \\ &Q_{concat} = (Q \oplus H_{enc}) W_q \\ &K_{concat} = (K \oplus H_{enc}) W_k \\ &\qquad V' = V W_v \end{aligned} \tag{1}$$

where $\oplus$ means concatenating two matrices in this equation.

Note that cross+self-attention, as well as standard self-attention, assume Q, K, and V to be identical matrices, namely, $Q = K = V = H_{dec}$, where H_{dec} is the matrix form of concatenating input vectors of the corresponding decoder layer.

10.5B Model As demonstrated in Kaplan et al. (2020), the performance of neural models improves as the number of parameters increases. Moreover, previous WMT shared tasks systems, such as Chen et al. (2020), achieved improvements in translation quality using model scaling. Following this insight, we attempted to scale up the translation model. Considering the constraints of GPU memory and training time, we finally configured the model size to be 10.5B parameters.

We also applied the position encoding methods used in last year's submission system (Morishita et al., 2022b). Namely, in the encoder, we employed relative position encoding (Shaw et al., 2018). In the decoder, we used SHAPE (Kiyono et al., 2021). We specified the maximum shift size of SHAPE to be 10.

Previous year's submission models We also incorporated the transformer models developed for the previous year's submission system as the primary translation module. We introduced the bottom-to-top (B2T) connection (Takase et al., 2023) to these models for training stability and relative position encodings (Shaw et al., 2018) to improve their generalization ability to unseen sentence lengths during training. For more details, please refer to (Morishita et al., 2022b).

4.2 Pre-Training

We trained each translation model shown in Table 4 with the filtered bitext and synthetic data described in Section 3.2. In this phase, we used the pre-training configuration shown in Table 1.

Following last year's submission system (Morishita et al., 2022b), the bitext was upsampled until it reached to a ratio of 1:1 with the synthetic data. Moreover, we used the tagged back-translation technique (Caswell et al., 2019) by adding a special token ⟨BT⟩ to the beginning of the source sentences in the synthetic data.

4.3 Fine-Tuning

The fine-tuning data are detailed in Section 3.3, and the hyperparameters utilized during training are as described in Table 1.

4.4 Ensemble

We ensembled the fine-tuned models, except for the 10.5B model, due to the computational resource limitations. We included the ensembled model and individual model outputs as the reranking candidates.

Direction	Configuration	#Params.	Cross+self attention	LN pos.	Encoder				Decoder			
					Layer	d_{model}	d_{ffn}	#Heads	Layer	d_{model}	d_{ffn}	#Heads
Both	NTT-Base	547M		Pre.	9	1024	8192	16	9	1024	8192	16
Both	ABCI-Base	622M		Pre.	9	1024	16384	16	9	1024	4096	16
Both	ABCI-EncBig	2.0B		Pre.	12	1024	65536	16	9	1024	8192	16
Both	ABCI-EncDeep	736M		Pre.	18	1024	8192	16	9	1024	8192	16
Both	Failab-EncBig	1.7B		Pre.	9	1024	61440	16	9	1024	16384	16
Both	Failab-DecBig	1.7B		Pre.	9	1024	16384	16	9	1024	61440	16
Both	NTT-A	408M		Post.	6	1024	8192	16	6	1024	8192	16
Both	NTT-B	547M		Post.	9	1024	8192	16	9	1024	8192	16
Both	NTT-C	622M		Post.	9	1024	16384	16	9	1024	4096	16
Both	NTT-D	698M		Post.	9	1024	16384	16	9	1024	8192	16
En-Ja	NTT-E	547M		Pre.	9	1024	8192	16	9	1024	8192	16
En-Ja	NTT-F	509M	✓	Post.	9	1024	8192	16	9	1024	8192	16
En-Ja	NTT-G	551M	✓	Post.	10	1024	8192	16	10	1024	8192	16
Both	Failab-LM	10.5B	✓	Pre.	16	4096	16384	32	32	4096	16384	32

Table 4: List of model configurations used by the primary translation module. The upper half of the table shows the models also used in last year's submission system (Morishita et al., 2022b), and the lower half shows the models newly trained this year. d_{model} and d_{ffn} respectively denote sizes of embedding and feedforward layers. LN pos. means the position of layer normalization. Post. denotes that layer normalization is applied after the residual connection. Pre. denotes that layer normalization is performed before the residual connection. ABCI-Base and NTT-Base were each trained with two different seeds.

5 Reranking

To enhance translation quality, we applied a reranking process to the candidate set of hypotheses translated by each model described in Section 4. We conducted a comparative analysis of the various methods, as presented in the following sections.

5.1 Methods

The reranking approach was used to obtain the final output $\hat{y}$ from C, where C represents the candidate set generated by multiple translation models for a given source x.

Quality Estimation (QE) This approach involves scoring the candidates using quality estimation methods (e.g., COMET-QE) and selecting the one with the highest score, as follows:

$$\hat{y} = \underset{c \in \mathcal{C}}{\operatorname{argmax}} \, \text{QE}\,(x, c)\,. \tag{2}$$

where, $\text{QE}(\cdot, \cdot)$ is a quality estimation function.

Minimum Bayes Risk (MBR) This method uses reference-based metrics such as COMET, to yield the best output as follows (Fernandes et al., 2022);

$$\hat{y} = \underset{c_i \in C}{\operatorname{argmax}} \frac{1}{|C|} \sum_{c_j=1}^{|C|} \text{RefMetric}\,(c_i, c_j)\,. \tag{3}$$

where $\text{RefMetric}(\cdot, \cdot)$[6] is a reference-based metric. Note that MBR uses reference-based metrics but not reference texts. MBR is applied to the output of a single model in Fernandes et al. (2022). We extended this method to the outputs from multiple models.

MBR after QE (QE → MBR) This approach is a combination of QE and MBR (Fernandes et al., 2022). We denoted the top-p samples from set C, according to the score calculated by the quality estimation function $\text{QE}(\cdot, \cdot)$, as $C_{\text{top-p}}$. Then, MBR was applied for $C_{\text{top-p}}$.

5.2 Post Evaluation

We experimented with the performance of the translation models and the reranking process. Note that this experiment was conducted after the primary system was submitted.

5.2.1 Experimental Setup

We used WMT21-COMET-QE[7] and WMT22-CometKiwi (Rei et al., 2022b)[8] for the QE, and WMT22-COMET-DA[9] as the refernece-

[6] Some reference-based metrics, such as COMET, also use source x as an input.

[7] https://unbabel-experimental-models.s3.amazonaws.com/comet/wmt21/wmt21-comet-qe-mqm.tar.gz

[8] https://huggingface.co/Unbabel/wmt22-cometkiwi-da

[9] https://huggingface.co/Unbabel/wmt22-comet-da

Models		En→ Ja	Ja → En
NT5	single model	8	8
	4-models ensemble	1	1
	all models ensemble	1	1
NTT	single model	70	40
	all models ensemble	10	10
Failab-LM		10	10
Total		100	70

Table 5: Breakdown of candidates for reranking. The NT5 four-model ensemble consists of ABCI-EncBig, ABCI-EncDeep, Failab-EncBig, and Failab-DecBig. The NT5 all-model ensemble consists of NTT-Base (two different seeds), ABCI-Base (two different seeds), ABCI-EncBig, ABCI-EncDeep, Failab-EncBig, and Failab-DecBig. The NTT all-model ensemble consists of NTT-A to NTT-G.

based metric for MBR. WMT22-COMET-DA was also used as the evaluation metric. The candidate sets contained 100 hypothesis for En→Ja and 70 for Ja→En. The breakdown of each candidate set is shown in Table 5.

5.2.2 Reranking Analysis

Table 6 shows the results of the reranking. Oracle (a) is the upper-bound setting, selecting the final output by using WMT22-COMET-DA with reference text (denoted r):

$$\hat{y} = \underset{c \in \mathcal{C}}{\operatorname{argmax}} \, \texttt{WMT22-COMET-DA}\,(c, r)\,. \quad (4)$$

Comparing the QE and MBR approaches (f and g vs. q) showed that MBR achieved higher performance. As for the QE approach, WMT21-COMET-QE achieved better performance than WMT22-CometKiwi in both translation directions (f vs. g). Therefore, we used WMT21-COMET-QE for the QE → MBR approach. The best performance was achieved by the QE → MBR at smaller p (h, i, j and k) in both translation directions. Moreover, QE → MBR often achieved a higher performance than MBR. These results suggest that the poor quality hypothesis in the candidates has a negative impact on MBR reranking.

5.2.3 10.5B Model Analysis

As described in Section 4.1, we trained a large-scale translation model with 10.5B parameters (failab-LM). The experimental results showed that the 10.5B parameters models were inferior to the best single model. However, when comparing the loss, we found that the 10.5B parameters models achieved a lower loss than the other smaller models. These results might suggest that 10.5B is overparametrized for sentence-level translation. For document-level translation, there may be an opportunity to harness the potential of the large number of parameters. However, the availability of document-level parallel corpora for En↔Ja is limited, highlighting the necessity of expanding the resources for document-level data.

In studies on large language models (LLMs), several papers discuss the scaling laws. For example, Hoffmann et al. (2022) introduces the optimal number of tokens with respect to model size, which is often referred to as the Chinchilla rule in the community. If we straightforwardly apply this rule to MT models, the optimal tokens of the 10B parameters MT model are estimated to be 205.1B tokens. This is much larger than the tokens we used to train for 10.5B parameter models. Therefore, we posit that effectively harnessing the 10.5B model may be possible by increasing both the quantity of training data and the number of training steps. We could not investigate this perspective due to the limited time and computational resources. Thus, we leave to clarify this perspective for future work.

5.2.4 Effectiveness of applying cross+self-attntion

In a preliminary experiment, we confirmed the effectiveness of applying cross+self-attention by comparing performance with the standard setup (cascading computation of self- and cross-attentions) of Transformer encoder-decoder models. Table 7 shows the results of our preliminary experiments. As we see, there were no considerable performance degradations when we compared the performance of cross+self-attention models (NTT-F) with those of standard self-attention and cross-attention cascading models (NTT-B).

In addition, cross+self-attention models reduce the computation of cascading self- and cross-attention into single cross+self-attention. Therefore, the cross+self-attention models are slightly faster and require less memory than standard self-attention and cross-attention cascading models.

6 Submission System

Initially, we planned to submit several versions of the system, with the highest-scoring system selected as the final version. However, the reranking process took longer than expected, and we were

ID	Candidates	Reranker	En→Ja		Ja→En	
			wmt22test	wmt23test	wmt22test	wmt23test
(a)	All	`Oracle`	0.9298	0.9136	0.8804	0.8737
(b)	Failab-LM	-	0.8840	0.8590	0.8127	0.8119
(c)	NT5-ensemble	-	0.8926	**0.8713**	**0.8269**	**0.8234**
(d)	NTT-ensemble	-	0.8880	0.8633	0.8215	0.8198
(e)	Best Single Model	-	**0.8937**	0.8692	0.8232	0.8198
(f)	All	`WMT21-COMET-QE`	**0.9085**	**0.8879**	**0.8379**	**0.8345**
(g)	All	`WMT22-CometKiwi`	0.9049	0.8847	0.8338	0.8329
(h)	All	`QE(Top10%)` → `MBR`	0.9102	**0.8905**	0.8425	0.8372
(i)	All	`QE(Top20%)` → `MBR`	**0.9111**	0.8904	**0.8437**	0.8393
(j)	All	`QE(Top30%)` → `MBR`	**0.9111**	**0.8905**	0.8425	0.8394
(k)	All	`QE(Top40%)` → `MBR`	0.9107	0.8903	0.8429	**0.8402**
(l)	All	`QE(Top50%)` → `MBR`	0.9099	0.8901	0.8431	0.8401
(m)	All	`QE(Top60%)` → `MBR`	0.9096	0.8897	0.8426	0.8401
(n)	All	`QE(Top70%)` → `MBR`	0.9092	0.8897	0.8418	0.8396
(o)	All	`QE(Top80%)` → `MBR`	0.9092	0.8892	0.8411	0.8390
(p)	All	`QE(Top90%)` → `MBR`	0.9088	0.8891	0.8408	0.8389
(q)	All	`MBR`	0.9084	0.8890	0.8405	0.8384

Table 6: Post evaluation results. Best Single model (b) represents the highest score achieved by an individual translation model (not an ensembled model).

Configuration	Cross+self attention	#Params.	En→Ja	
			wmt22test	wmt23test
`NTT-B`		547M	0.8865	0.8624
`NTT-F`	✓	509M	0.8862	0.8612
`NTT-G`	✓	551M	0.8862	0.8635

Table 7: Comparison of performance on applying cross+self-attention compared with the standard setup (cascading computation of self- and cross-attentions) of Transformer encoder-decoder models.

unable to submit multiple submissions within the time limit. Therefore, the system that was actually submitted system was slightly different from the one described in this paper, as follows:

- For the En→Ja system, we submitted the results of the ensembled model of NTT-A to NTT-G.
- For the Ja→En system, we opted for the `QE(Top 80%)` → `MBR` configuration.

Unlike the post evaluation setting (Section 3.3), these models were fine-tuned using all of the WMT'20 test set, the WMT'20 development set, the WMT'21 test set, and the WMT'22 test set.

7 Conclusion

This paper described our submission system for the constrained track of the WMT'23 general translation task. We developed a translation system for En↔Ja. We perform reranking on the candidates generated by multiple translation models, which include a large-scale model with 10.5 billion parameters. Post evaluation (Section 5.2) confirmed the limitations of sentence-level translation quality improvement through model scaling and the effectiveness of our reranking approach.

Acknowledgments

We thank an anonymous reviewer who provided feedback. Also, we would like to also appreciate the member of Tohoku NLP Group for their cooperation in conducting this research. This work was mainly done under the NTT-Tohoku University collaborative research agreement. The work of Jun Suzuki was partly supported by JST Moonshot R&D Grant Number JPMJMS2011 (fundamental research).

References

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged Back-Translation. In *Proceedings of the Fourth Conference on Machine Translation (WMT)*, pages 53–63, Florence, Italy. Association for Computational Linguistics.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. WIT3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy. European Association for Machine Translation.

Peng-Jen Chen, Ann Lee, Changhan Wang, Naman

Goyal, Angela Fan, Mary Williamson, and Jiatao Gu. 2020. Facebook AI's WMT20 News Translation Task Submission. In *Proceedings of the Fifth Conference on Machine Translation*, pages 113–125, Online. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:1706.02677*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *CoRR*, abs/2001.08361.

Shun Kiyono, Sosuke Kobayashi, Jun Suzuki, and Kentaro Inui. 2021. SHAPE: Shifted Absolute Position Embedding for Transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3309–3321.

Hayato Kobayashi. 2018. Frustratingly Easy Model Ensemble for Abstractive Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4165–4176, Brussels, Belgium. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and Maja Popović. 2023. Findings of the 2023 Conference on Machine Translation (WMT23). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Taku Kudo. 2006. MeCab: yet another part-of-speech and morphological analyzer. http://mecab.sourceforge.net.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 66–71. Association for Computational Linguistics.

Makoto Morishita, Katsuki Chousa, Jun Suzuki, and Masaaki Nagata. 2022a. JParaCrawl v3.0: A Large-scale English-Japanese Parallel Corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6704–6710, Marseille, France. European Language Resources Association.

Makoto Morishita, Keito Kudo, Yui Oka, Katsuki Chousa, Shun Kiyono, Sho Takase, and Jun Suzuki. 2022b. NT5 at WMT 2022 General Translation Task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 318–325, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Graham Neubig. 2011. The Kyoto Free Translation Task. http://www.phontron.com/kftt.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Stephan Peitz, Sarthak Garg, Udhay Nallasamy, and Matthias Paulik. 2019. Cross+self-attention for transformer models. `https://github.com/pytorch/fairseq/files/3561282/paper.pdf`.

Reid Pryzant, Youngjoo Chung, Dan Jurafsky, and Denny Britz. 2018. JESC: Japanese-English Subtitle Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are References Really Needed? Unbabel-IST 2021 Submission for the Metrics Shared Task.

In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 86–96. Association for Computational Linguistics.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-Attention with Relative Position Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 464–468.

Sho Takase, Shun Kiyono, Sosuke Kobayashi, and Jun Suzuki. 2023. B2T Connection: Serving Stability and Performance in Deep Transformers. In *Findings of the Association for Computational Linguistics (ACL)*, pages 3078–3095, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *CoRR*, abs/2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems 31 (NIPS)*, pages 5998–6008.

KYB General Machine Translation Systems for WMT23

Shivam Kalkar, Ben Li and Yoko Matsuzaki

NRI Digital, Ltd,
{s-kalkar, b4-li, y-matsuzaki}@nri.co.jp

Abstract

This paper describes our approach to constructing a neural machine translation system for the WMT 2023 general machine translation shared task. Our model is based on the Transformer architecture's base settings. We optimize system performance through various strategies. Enhancing our model's capabilities involves fine-tuning the pretrained model with an extended dataset. To further elevate translation quality, specialized pre- and post-processing techniques are deployed. Our central focus is on efficient model training, aiming for exceptional accuracy through the synergy of a compact model and curated data. We also performed ensembling augmented by N-best ranking, for both directions of English to Japanese and Japanese to English translation.

1 Introduction

In the context of the WMT 2023 general machine translation shared task for Japanese to/from English, we tackle the inherent challenges posed by the diverse linguistic structures of these languages. The transformative impact of the Transformer model on neural machine translation is undeniable. While current trends prioritize larger models and extensive datasets, our focus remains on achieving efficient translation with modest resources. This study underscores our use of a compact model and limited computational assets to enhance translation quality.

Built upon the Transformer model's base settings, our approach uses pre-trained models trained on Japanese-English parallel data (Morishita et al., 2019). A previous study involved fine-tuning on various datasets, yielding excellent translation within specific domains (Kalker et al., 2021). In this study, we refined our fine-tuning dataset and systematically tuned hyperparameters to optimize results. Post-fine-tuning, we harnessed model ensembling techniques to amalgamate multiple model outputs, leading to better translation quality. Our study highlights the specifics of our system configurations and methods, offering a concise overview of our strategies.

2 Data selection and Preprocessing

In this section, we elaborate on the process of creating our fine-tuning dataset for the Neural Machine Translation (NMT) system, with a focus on enhancing translation quality for the WMT competition. Our approach involved meticulous data selection and preprocessing to ensure the effectiveness of our system. We describe the details behind selecting the base dataset, incorporating additional parallel corpora, and performing data cleaning to curate a high-quality training dataset.

2.1 Base Dataset Selection

Our foundational dataset for training the initial NMT models is derived from the JParacrawl Version 3 dataset, which offers a diverse array of content spanning various domains. This choice was made due to its comprehensive coverage, which provides a strong starting point for training the base NMT models.

2.2 Augmenting the Dataset

Upon training our base models, we identified an opportunity to further enhance translation quality by incorporating additional datasets. To achieve this, we integrated parallel corpora obtained from sources recommended by the WMT competition organizers. This augmentation was aimed at increasing the diversity of the training data, which often contributes to improved translation accuracy.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 137–142
December 6–7, 2023.

2.3 Data Cleaning

Data cleaning played a pivotal role in refining the quality of our training dataset. During this stage, we implemented several key steps to ensure the integrity of the data:

Language Focus: Given our goal of improving Japanese-to-English translation, we focused on maintaining language homogeneity within the dataset. Therefore, non-Japanese languages, such as Korean and Chinese, as well as their corresponding English translations, were deleted from the dataset.

Sentence Length and Quality: To uphold the overall coherence and effectiveness of the NMT model, we eliminated sentences that were excessively short or of low quality. This step aimed to prevent the model from learning suboptimal translation patterns and to maintain a high standard of translation output. Furthermore, to maintain coherence, we curate our training data by excluding sentences longer than 250 subwords (see 3.1).

Translation Pair Quality: Translation Pair Quality: JParacrawl v3 provides a score for translations labeled as "Accuracy", so we removed sentences with lower scores from the dataset. The threshold for the score was set at 0.5 for training data and 0.75 for the validation dataset. This procedure was not applied to the other parallel corpora.

Normalize symbolic characters: Normalize symbolic characters: Especially for Japanese sentences, since the language has more variations of symbolic characters like 「」, 『』, and "" for quotation marks. We added pre-processing to normalize symbolic characters based on rules. We decided that emojis were not included in this process, as these characters are translated as they are. By adding these rule-based translations, the final BLEU score increased by +0.1.

2.4 Final Dataset Composition

The outcome of our data selection and preprocessing efforts yielded a curated dataset (22.2M). We divided the processed data into a training dataset and a test dataset (2.5M) to evaluate model quality. The training dataset was further divided into a train and a validation to perform fine-tuning on our NMT models.

We also developed a fusion dataset by combining the processed JParacrawl v3 training data with other parallel corpora that were provided by WMT 2023. The dataset finally contains 49.9M sentences (Figure 1, Table 1).

Figure 1 Dataset development

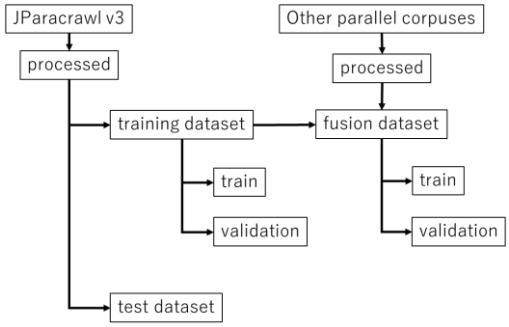


Table 1 Data selection summary

Dataset	Sentences
JParaCrawl Ver.3	25.7M
JParaCrawl Ver.3 (processed)	22.2M
Other Parallel Corpus	33.8M
Other Parallel Corpus (processed)	27.7M
Fusion Corpus (JParacrawlVer3 + Other Parallel)	49.9M

3 Tokenization

3.1 SentencePiece Toolkit for Tokenization and Detokenization

We use the SentencePiece toolkit (Kudo and Richardson, 2018) for tokenization. SentencePiece is suited for languages with complex linguistic structures and compound words. Its efficacy is pronounced in languages with ambiguous word boundaries, agglutinative morphology, and compound word usage. This enables the extraction of subword components from intricate terms, enhancing our tokenization precision. It can remove meta-symbols from translated output, ensuring fluidity and linguistic correctness in the final translations.

3.2 Customized Vocabulary

To bridge vocabulary disparities between our base model (JParacrawl Version 1) and our Fusion Corpus, we train a SentencePiece tokenizer. This

tokenizer aligns with our data's linguistic nuances, enhancing token accuracy. Our SentencePiece model employs a vocabulary size of 32,000 tokens.

4 Model Training

This section details the training process of our translation models using the fairseq toolkit. The selection and configuration of models, as well as the optimization parameters, are presented. Additionally, our model training strategy (Figure 2), including the utilization of advanced techniques such as mixed-precision training and beam search during decoding, is outlined.

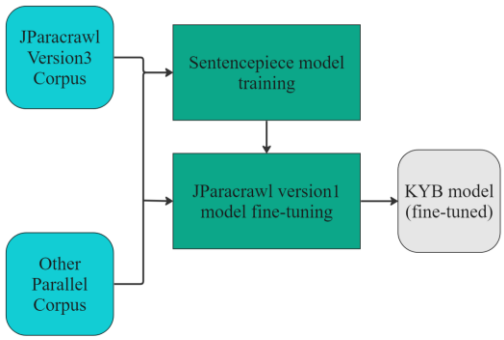
4.1 Model Selection and Configuration

From the array of available models, including MBART and JParacrawl, our evaluation led us to opt for Jparacrawl due to its favorable accuracy-performance trade-off. Jparacrawl models are underpinned by the Transformer architecture (Vaswani et al., 2017) with base settings. The encoder and decoder feature six layers each, embedding sizes of 512, and feed-forward embedding sizes of 2048. Eight attention heads are employed for both the encoder and decoder. Dropout with a probability of 0.3 is applied to enhance generalization. The Adam optimizer with $\alpha = 0.001$, $\beta1 = 0.9$, and $\beta2 = 0.98$ is utilized. A square root decay learning rate schedule with a linear warmup of 4000 steps is implemented. Gradient clipping maintains stability by ensuring gradients do not exceed a norm of 1.0. Mini-batches contain around 5,000 tokens, with gradient accumulation of 64 mini-batches per update. Training spans 24,000 iterations, with model parameter snapshots saved every 200 iterations. The final model is an average of the last eight snapshots. The use of mixed-precision training optimizes performance on modern GPUs.

4.2 Decoding Strategy

During decoding, various beam search sizes (2, 4, 6, 8) were employed to compare translation results. We also compared the best checkpoint and averaged checkpoint to obtain various translation results. In the training of the Ja to En model, we also compared the optimal trade-off between translation quality and computational efficiency by optimizing the training precision.

Figure 2 Model training system



4.3 Training Environment

Our model training is executed on Google Cloud Platform's compute engine equipped with 4-T4 GPUs. Mixed precision (float16) training takes approximately 12 hours and full precision (float32) training takes approximately 30 hours. Although the BLEU score is higher for full precision, the difference is not so large (+0.6, Table 2). So, we decided to use mixed precision for training the Fusion Corpus. We used the train-validation-test split ratio of 90:5:5.

Table 2 En-Ja models summary

Test Dataset	Model	Training Precision	BLEU score	change
JPC V3	JPC V1	-	38.0	0.0
JPC V3	KYB	Mixed (fp16)	45.3	+7.3
JPC V3	KYB	Full (fp32)	45.9	+7.9
Fusion	JPC V1	-	22.3	0.0
Fusion	KYB	Mixed (fp16)	29.8	+7.5

※JPC: Shorthand for JParacrawl
*The results used the best checkpoint with beam size 4. Test set was based on JPC V3 data.

Table 3 Ja-En models summary

Test Dataset	Model	Training Precision	BLEU score	change
JPC V3	JPC V1	-	19.2	
JPC V3	KYB	Full (fp32)	43.8	+24.6

*The results used the best checkpoint with beam size 4. Test set was based on JPC V3 data.

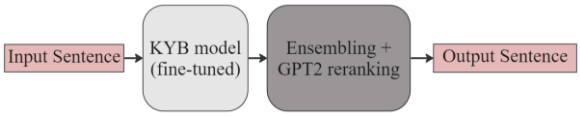
4.4 Performance Evaluation

To assess the efficacy of our trained models, we employ validation data from our dataset. The sacreBLEU metric (Post, 2018) is employed to calculate BLEU scores, offering a quantitative evaluation of translation quality. For testing, we used a set of 2.5M sentences from JParacrawl v3.

5 Model Ensembling with N-Based Reranking

In this section, we delve into the intricacies of our advanced model ensembling approach (Figure 3) coupled with N-Based Reranking, a technical strategy inspired by Le et al. (2021). Our objective is to optimize translation quality through a combination of models and a refined reranking mechanism.

Figure 3 Model inference system



5.1 Model Averaging via Ensembling

Our ensembling technique involves the aggregation of multiple trained model files. Through model averaging, we synthesize the insights and strengths of various models into a unified translation framework. This process not only enhances the stability of our translation outputs but also contributes to an overall improvement in translation quality. Moreover, we implemented checkpoint averaging by considering the last eight checkpoints to create an averaged model and employed in-training evaluation to identify the best checkpoint model.

5.2 N-Based Reranking Strategy

The crux of N-Based Reranking revolves around the calculation of token probabilities and sentence perplexities for translations generated by distinct checkpoint files of our fine-tuned model. We generate 4 alternative translations for each source sentence by using different beam search sizes for one checkpoint file. We compared 6 checkpoint files for the En-Ja side from three different trained models, which yielded 24 different translations for each source sentence. For the Ja-En side, we use two different checkpoints from trained models and use the previous study's model to make 12 alternative translations for each source sentence. This multifold approach introduces diversity into our pool of translation candidates, a crucial aspect of refining translation quality.

To identify the optimal translation candidate among these alternatives, we employ a GPT-2 based ranker (Radford et al., 2019). The ranker computes perplexity for each alternative translation, then chooses the lowest perplexity score as the most proper translation. We submitted the best translation result from all the alternative translations.

6 Post-processing of translation

We found specific tendencies of mistranslation, such as adding double quotation marks, deleting part of the quotation marks, or repeating a specific word endlessly in our translation results. These phenomena are typical problems in machine translation tasks, so we added post-processing to reduce the mistranslations.

After the post-processing, we submitted our results. Our final submission was scored as shown in Table 4 in the automatic evaluation.

Table 4 Submission results

Submission	COMET	BLEU	chrF
Ja-En	76.6	17.6	43.9
En-Ja	80.8	17.8	27.7

7 Discussion and Future Work

7.1 Discussion

In this study, we participated in the general translation task to achieve better translation quality with a relatively compact model and dataset. We made two different datasets: one is based on JparaCrawl version 3 data (JPC V3), and the other included additional parallel corpora provided by WMT23 (Fusion Corpus). The results of our local test suggested that the JPC V3 fine-tuned model shows a better BLEU score than the Fusion Corpus fine-tuned model. However, we exercised caution in interpreting these scores, since the test dataset

contains only sentences from Jparacrawl v3 dataset. The model trained with same origin data might show the result of overfitting to the domain of JPC3 dataset, even the test and train dataset contains different sentences. We then tried to ensemble these different models to achieve more robust translation results. Additionally, we generated alternative translations using different inference parameters for each model and then chose the most proper translation by using PPL.

We use GPT-2 to calculate PPL in this study since the model's knowledge of language is somewhat better than BERT's (see Appendix for a comparison of PPLs). Using a pre-trained large model is obviously effective when computational resources are limited. We use PPL to select a better result from the alternative translations; however, PPL is a relative metric of how fluent the sentence is or how acceptable the sentence order is for the model, so it is not a direct metric of translation quality (Kalkar et al., 2022; Wang et al., 2022). According to that, we recognize that the method that relies on PPL still has some limitations in improving the quality of translations.

Additionally, we added some post-processing to reduce specific mistranslations through the model. Although we carefully cleaned up noisy symbolic characters such as quotation marks from the training dataset, the model's output is still not reliable for translating symbolic characters properly. We performed some rule-based translations to modify symbolic characters.

7.2 Future Work

Back-Translation / Forward-Translation

To improve our translation pipeline, we explored the integration of back-translation as a potential enhancement. Back-translation involves using a trained model to translate from the target language back to the source language (forward-translation is vice versa), effectively creating a synthetic parallel dataset. While we attempt to do the back-translation, we need to consider the quality of the synthetic dataset, especially the variety of translations. When our synthetic dataset does not have enough variation in translation, the model can be easily overfit to the specific translation pattern.

To avoid that, we tried to use a common API for translation, like the Google API; however, this proved to be a very time-consuming task to obtain enough dataset for training (e.g., when one response takes 1 sec, it takes more than 55h to obtain 200,000 sentences). In this study, we attempted to perform back/forward translation, however, we were not able to obtain enough volume of content with reasonable quality. We would like to find a practical method to develop datasets with reasonable quality for back or forward translation in future work.

8 Conclusion

In this study, we embarked on an extensive exploration of high-efficiency model training strategies, leveraging limited computational resources alongside a streamlined model architecture rooted in the Transformer framework's base settings. Our investigation yielded crucial insights and techniques that converge to create a high-quality translation system. Through our experimentation, we identified data cleaning, model averaging, ensembling, beam search, finetuning, parameter-tuning, and post-processing as pivotal techniques, enhancing the quality of our compact model and modest dataset.

References

1. Makoto Morishita, Jun Suzuki and Masaaki Nagata. 2019. JParaCrawl: A large scale web-based English-Japanese parallel corpus. *arXiv preprint* arXiv:1911.10668. https://doi.org/10.48550/arXiv.1911.10668

2. Shivam Kalkar, Yoko Matsuzaki, and Ben Li. 2022. KYB General Machine Translation Systems for WMT22. https://aclanthology.org/2022.wmt-1.22/

3. Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics. https://aclanthology.org/D18-2012

4. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser and Illia

Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30. https://doi.org/10.48550/arXiv.1706.03762

5. Giang Le, Shinka Mori, and Lane Schwartz. 2021. Illinois Japanese↔ English News Translation for WMT 2021. In *Proceedings of the Sixth Conference on Machine Translation*, pages 144–153, Online. Association for Computational Linguistics.. https://aclanthology.org/2021.wmt-1.11

6. Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics. https://aclanthology.org/W18-6319

7. Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

8. Yequan Wang, Jiawen Deng, Aixin Sun, and Xuying Meng. 2022. Perplexity from PLM Is Unreliable for Evaluating Text Quality. https://doi.org/10.48550/arXiv.2210.05892

Appendix-1: Perplexity comparison BERT vs GPT2 for Japanese

PPL metric: Lower PPL corresponds to better semantic quality.

Example1: BERT gives low score for bad sentences

Sentence	BERT	GPT2
返品は与えられたものではありません! [Returns are not allowed!]	15.9	26.3
NATO ストラップは、時計の下にループし、最後に追加のキーパーを通します。 [The NATO strap loops under the watch and finally passes through an additional keeper.]	9.7	45.6
この音は、マスターユニットによるセカンドロックです。 [This sound is of the second lock by the master unit.]	12.5	33.4

Example2: BERT gives high score for good sentences

Sentence	BERT	GPT2
国際郵便 - 日本郵便 [International Mail - Japan Post.]	49.6	1.9
大切なことは、毎晩 3 つのことを書き続けることです。 [The important thing is to keep writing three things every night.]	23.5	8.7
タングステン重合金は無毒で環境にも優しいため、子供や大人がタングステン重合金を扱ったり作業したりするのに安全です。 [Tungsten heavy alloys are non-toxic and environmentally friendly, making it safe for children and adults to handle or work with tungsten heavy alloys.]	78.2	6.8

Yishu: Yishu At WMT2023 Translation Task

Qiulin Chen 415626524@qq.com
Dtranx AI, Chengdu, China

Min Luo 1148859199@qq.com
Dtranx AI, Chengdu, China

Yixin Tan 420129823@qq.com
Dtranx AI, Chengdu, China

Abstract

This paper introduces the Dtranx AI translation system, developed for the WMT 2023 Universal Translation Shared Task. Our team participated in two language directions: English to Chinese and Chinese to English. Our primary focus was on enhancing the effectiveness of the Chinese-to-English model through the implementation of bilingual models. Our approach involved various techniques such as data corpus filtering, model size scaling, sparse expert models (especially the Transformer model with adapters), large-scale back-translation, and language model reordering. According to automatic evaluation, our system secured the first place in the English-to-Chinese category and the second place in the Chinese-to-English category.

1 Introduction

This year, the Dtranx AI team participated in the WMT2023 Universal Translation Sharing Task and focused on enhancing the performance in the zh-en and en-zh language directions. For data preprocessing, we employed various methods, including knowledge-based rules, language detection, and language modeling, to clean parallel, monolingual, and back-translated data. Our data primarily comprised large-scale data mining and back-translation. Additionally, we applied punctuation regularization, byte pair encoding (BPE) Sennrich et al. (2015), and subword regularization methods Provilkov et al. (2019) for processing the data, which yielded excellent results across all languages.

In the modeling section, we have made enhancements to Fairseq Ott et al. (2019) by increasing the model's depth and width. Specifically, we augmented the Transformer model et al.Vaswani et al. (2017) by significantly increasing the number of layers and widening the model architecture.This modification allows the model to capture more complex patterns and dependencies in the data. Additionally, we have embraced the concept introduced by Bapna et al. Bapna et al. (2019) to expand the Transformer model by incorporating language-specific adapters, thus bridging the gap between diverse languages. Lastly, we integrated the dense Transformer model with the sparse Adapter model and leveraged the language model to reranking the final results, leading to further improvements in system performance.

For both English and Chinese translation tasks, we have developed separate systems. We have enhanced the model capacity and applied Adapter fine-tuning techniques, while also incorporating additional proprietary data for system training purposes.During the model inference phase, we have implemented reordering techniques to select more optimal translations. Based on automatic evaluation, our system achieved first place in English to Chinese translation and second place in Chinese to English translation.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 143–149
December 6–7, 2023.

2 Method

2.1 Data

In this section, we will present our primary dataset, which consists of bitext data and monolingual data sources, along with the preprocessing methods employed to prepare this initial data. Additionally, we will provide details about the setup utilized for training our baseline model.

2.1.1 Bitext Data

For the Chinese-English-English-Chinese language pairs, we utilize all bitext data in the shared task and include additional data sources for English-Chinese conversion. During data processing, we implemented the following knowledge-based rules for enhancement:

Remove empty sentences.
Eliminate escaped HTML characters.
Standardize different punctuation variations.
Normalize spacing.
Remove sentences with repetition marks, including single characters repeated more than four times, two characters repeated more than three times, and three characters repeated more than twice.
Delete sentence pairs with inconsistent punctuation at the end of the original and translated texts.
Remove sentence pairs with a source/target token ratio exceeding 1:3 (or 3:1).
Delete segments that exceed 150 tokens in length.
Remove sentence pairs with fewer than 5 tokens in the source text or translation.
Convert traditional Chinese characters to simplified Chinese characters.
Delete corpora with an unaligned number of parentheses.
Delete corpora with an unaligned number of Arabic numerals.
Remove corpora with non-native character ratios greater than 0.4.

We employed Moses Koehn et al. (2007) for normalizing spacing and punctuation. We utilized all accessible data sources to train our model.

Considering the aforementioned concerns regarding corpus quality, we implemented additional filtering steps to ensure data availability. Initially, we attempted to filter out low-quality sentence pairs using the word alignment method of fast-align Dyer et al. (2013). We retained the top 80% of sentence pairs based on the alignment score(a score generated by the word alignment model that measures the quality of word alignment between source and target sentences), encompassing all directions. Subsequently, we trained the Transformer model for all languages using Fairseq, following a similar approach as outlined in the study conducted by Bei et al. Bei et al. (2019). The scores were calculated as follows:

$$Score_{sentence} = PPL \tag{1}$$

$$Score_{combine} = \lambda * Score_{src} + (1 - \lambda) * Score_{tgt} \tag{2}$$

Here, we employ PPL as an abbreviation for perplexity, which represents the perplexity of the sentence language model. The value of λ, on the other hand, is determined empirically

based on language pairs and ranges from 0.2 to 0.8. For instance, if our source language is English and the target language is Chinese, we would set λ to 0.7.

Finally, the training data, as presented in Table 1, was carefully curated to serve as the foundational resource for our model training. This bilingual training dataset consists of 50 million sentence pairs in the Chinese-English (zh-en) language pair,and it can be utilized bidirectionally.

Language Pair	Data
zh-en	50M

Table 1: Ultimate bitext training data

Language Pair	Data
zh	72M
en	10M

Table 2: Ultimate monolingual data

2.1.2 Monolingual Data

To ensure the quality of our data and to create synthetic parallel texts, we harnessed the capabilities of a well-trained bilingual model. We compiled high-quality monolingual corpora in various languages from reputable sources, including news commentaries, europarl, and news crawls. The monolingual data, after undergoing a rigorous filtering process, is presented in Table 2.

2.1.3 Tokenizer

We opted for SentencePiece Kudo and Richardson (2018) as the training tool for our subword tagger. To enhance subwording efficiency, we adopted the approach of Tran et al. Tran et al. (2021) by employing sampled text with a temperature of 5. For the bilingual model, we utilized a vocabulary of 32,000 words.

In addition, we integrated the subword regularization method Provilkov et al. (2019) Raffel et al. (2020) into the tagged text. This technique was exclusively applied to the source side, as it has the potential to enhance the model's robustness by allowing different subword tokenizations.

2.2 Model Architectures

We have developed a dedicated model for bidirectional translation between Chinese and English, capitalizing on our proficiency as native Chinese speakers. To enhance the quality of our training data, we integrated a private corpus comprising approximately 20 million high-quality sentence pairs spanning various domains, including general text, technology, medicine, law, finance, and more.

Regarding "basic fine-tuning," our approach involves parameter adjustments, including fine-tuning the learning rate, the number of training epochs, and batch sizes. These adjustments are made to optimize the model's adaptation to the specific translation task at hand.

In the realm of back-translation, we harness English and Chinese monolingual corpora. This approach leverages monolingual text data in both languages, enriching the model's translations by back-translating them into English. This technique seamlessly augments the diversity of our training dataset.

As for the specifications of our Transformer model, it features 12 encoder layers and 6 decoder layers, each equipped with 8 attention heads. The embedding size is set to 512, and the width of the feed-forward neural network (FFN) is 4096. Additionally, we have incorporated techniques like Layer Normalization and residual connections to stabilize the model training process.

2.2.1 Language Specific Adapter

In essence, a language-specific adapter layer is a dense layer that incorporates residual connections and nonlinear projections. The hyperparameter "b" represents the dimension of the internal dense layer. These adapter layers consist of a multitude of globally shared parameters, along with several task-specific layers. This unique design allows us to train and optimize individual models for multiple languages.

Bapna et al. Bapna et al. (2019) demonstrated the improved translation performance achieved by machine translation models employing adapters. Therefore, following the training of our bilingual models, we integrated adapter layers into them and subsequently conducted additional training and fine-tuning on these adapter layers. To be specific, for the Chinese-English and English-Chinese models, we introduced a language-specific adapter with a dense layer dimension of 4096.

Regarding the incorporation of adapters in the bilingual models, we seamlessly integrate the adapter layers into the existing architecture, where they operate alongside the standard layers. The globally shared parameters refer to the model parameters that are common across various languages and tasks, which are shared among different adapter layers in the model.

For the fine-tuning of the adapters, we utilized additional bilingual data specific to the translation tasks. These adapters were fine-tuned with the same data used for training the main translation model, allowing them to adapt to the particular translation requirements of our task.

2.2.2 Finetune

To enhance the model's performance, we implemented in-domain fine-tuning, a proven effective technique in previous news translation tasks. We generated various types of fine-tuned data using the following approach. According to the studies conducted by Li et al. Li et al. (2020) and Wang et al.Wang et al. (2021), low-frequency and high-frequency words often pertain to domain-specific nouns and other related terms that directly reflect the topic at hand. However, this year's shared task has transitioned from the news domain to a more generalized translation task. Recognizing that previous fine-tuning using news domain data could potentially have a detrimental effect on the model, we adopted the strategy outlined by Li et al. Li et al. (2020) and Wang et al. Wang et al. (2021), which involves selecting topic-related data based on a test set. Subsequently, we identified specific data for further fine-tuning and conducted experiments on the 2022 news development set, subsequently applying the refined model directly to the 2022 test set.We fine-tuned the full model, not just the adapters, to ensure it was well-suited for the task at hand.

2.2.3 Model Ensemble

Model integration has been widely adopted as a technique in previous WMT sharing tasks. To mitigate bias towards more recent training data, it is common practice to average multiple checkpoint parameters of the model. Specifically, during training, we consistently take the average of the last five checkpoints. In the fine-tuning phase, we fine-tune the hyperparameters (e.g., num epoch and num average checkpoints) based on the performance on the development set and directly apply them to the test set of WMT23.

3 Results

3.1 Experimental setup

Each model was trained on eight NVIDIA A100 GPUs, each equipped with 40 GB of memory. Additionally, we employed high-volume processing and higher learning rates, as mentioned in Ott et al. Ott et al. (2018). The maximum learning rate was set to 0.0005, and we used 10,000 warm-up steps. All dropout probabilities were set to 0.1. To expedite training, we utilized half-precision floating-point numbers (FP16). In the context of multilingual training, we incorporated source language labels and target language labels to leverage the distinctions between languages. Following the approach proposed by Tran et al. Tran et al. (2021), we segmented the data into multiple parts and downsized the data in both the high-resource direction and in synthetic backtranslation for each training cycle.

3.2 Results

We trained bilingual models for English to Chinese (en-zh) and Chinese to English (zh-en). In our model, we enhanced the model capacity by introducing specific adapter layers for each translation direction, addressing the unique linguistic challenges of each language pair. These adapter layers do not induce sparsity; instead, they add more trainable parameters to the model. Each model has dedicated adapters for en-zh and zh-en, as they are not shared between the bilingual models. To refine our training set, we extracted additional relevant corpus from the raw text in the test set. This extracted data was structured using the language model and augmented through reverse translation. The results demonstrate the effectiveness of our systematic approach, and we achieved the highest scores on the COMET evaluation metric. These outcomes are detailed in Table 3 and Table 4

Team	Bleu	Chrf	Comet
HW-TSC	33.6	57.5	82.8
Yishu	33.4	57.4	82.7
GPT4-5shot	26.8	53.1	81.6
ZengHuiMT	27.0	54.6	79.6

Table 3: Submission results for zh-en in WMT23

Team	Bleu	Chrf	Comet
HW-TSC	58.6	53.8	87.3
Yishu	57.6	53.0	88.1
GPT4-5shot	49.6	46.5	87.1
ZengHuiMT	52.9	47.0	84.3

Table 4: Submission results for en-zh in WMT23

4 Conclusion

In this paper, we present Dtranx AI's submission to the WMT2023 Universal Translation Shared Task. For the Chinese-English and English-Chinese language pairs, we adopt a bilingual model as the fundamental structure and enhance it through various strategies. These include increasing the model capacity, fine-tuning with Adapters, incorporating private relevant corpus,

and optimizing the translation output by reordering. Our experimental results demonstrate the effectiveness of these optimization techniques.

References

Bapna, A., Arivazhagan, N., and Firat, O. (2019). Simple, scalable adaptation for neural machine translation. *arXiv preprint arXiv:1909.08478*. 1, 2.2.1

Bei, C., Zong, H., Yuan, C., Liu, Q., and Fan, B. (2019). Gtcom neural machine translation systems for wmt19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 116–121. 2.1.1

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648. 2.1.1

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics. 2.1.1

Kudo, T. and Richardson, J. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*. 2.1.3

Li, Z., Zhao, H., Wang, R., Chen, K., Utiyama, M., and Sumita, E. (2020). Sjtu-nict's supervised and unsupervised neural machine translation systems for the wmt20 news translation task. *arXiv preprint arXiv:2010.05122*. 2.2.2

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*. 1

Ott, M., Edunov, S., Grangier, D., and Auli, M. (2018). Scaling neural machine translation. 3.1

Provilkov, I., Emelianenko, D., and Voita, E. (2019). Bpe-dropout: Simple and effective subword regularization. *arXiv preprint arXiv:1910.13267*. 1, 2.1.3

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551. 2.1.3

Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*. 1

Tran, C., Bhosale, S., Cross, J., Koehn, P., Edunov, S., and Fan, A. (2021). Facebook ai wmt21 news translation task submission. arxiv. 2.1.3, 3.1

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30. 1

Wang, L., Li, M., Liu, F., Shi, S., Tu, Z., Wang, X., Wu, S., Zeng, J., and Zhang, W. (2021). Tencent translation system for the wmt21 news translation task. In *Proceedings of the sixth conference on machine translation*, pages 216–224. 2.2.2

PROMT Systems for WMT23 Shared General Translation Task

Alexander Molchanov & Vladislav Kovalenko
PROMT LLC
17E Uralskaya str. building 3, 199155,
St. Petersburg, Russia
First.Last@promt.ru

Abstract

This paper describes the PROMT submissions for the WMT23 Shared General Translation Task. This year we participated in two directions of the Shared Translation Task: English to Russian and Russian to English. Our models are trained with the MarianNMT toolkit using the transformer-big configuration. We use BPE for text encoding, both models are unconstrained. We achieve competitive results according to automatic metrics in both directions.

1 Introduction

The WMT Shared General Translation Task is an annual event where different companies and researchers build and test their systems on the test sets provided by the organizers. This year we decided to participate in two directions: English to Russian and Russian to English. We use the standard transformer-big configuration for our models. The English-Russian model is basically the same as last year, whereas the Russian-English model is a new one built for WMT23.

The rest of the paper is organized as follows: in Section 2 we describe in detail the systems we submitted to the Shared Task. In Section 3 we present and discuss the results. We conclude the paper in Section 4 with discussion for possible future work.

2 Systems overview

All of our WMT22 submissions are `MarianNMT`-trained (Junczys-Dowmunt et al., 2018) transformer-big (Vaswani et al., 2017) systems. We use the `OpenNMT` toolkit (Klein et al., 2017) version of byte pair encoding (BPE) (Sennrich et al., 2016b) for subword segmentation. Our BPE models are case-insensitive, we use special tokens in the source and target sides to process case (see Molchanov (2019) for details).

All of the systems are unconstrained, i.e. we use all data provided by the WMT organizers, all publicly available data and some private data crawled from different web-sources.

We also augment our training data with two types of synthetic data: 1) back-translations (Sennrich et al., 2016a) and 2) synthetic data with placeholders as described in Pinnis et al. (2017). The back-translations are obtained using the previous versions of our NMT models which are baseline transformers trained with less data (and without some up-to-date data like the news 2021 corpora from statmt.org). We also tag all our synthetic data with special tokens at the beginning of the source sentences as described in Caswell et al. (2019).

All models are trained with guided alignment which is used at translation time to handle named entities and document formatting. We obtain alignments using the `fast-align` (Dyer et al., 2013) tool.

The data statistics for the Russian-English language pair are presented in Table 1.

The details regarding different directions can be found in the next Section.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 150–154
December 6–7, 2023.

	Russian-English	
	#sent	#tokens RU
WMT+OPUS	37.4	690.9
Private	30.2	542.2
Total	67.6	1233.1

Table 1: Statistics for the filtered human parallel data in millions of sentences (#sent) and tokens (#tokens) for the English-Russian language pair. WMT stands for the data available for the News Task on the statmt.org/wmt22 website; OPUS is the data from the OPUS website apart from the data available for the News Task; Private stands for private company data.

2.1 Data preparation

There are several stages in our data preparation pipeline. These are mostly common filtering techniques. The main stages of the pipeline are:

- Basic filtering
  This includes some simple length-based and source-target length ratio-based heuristics, removing tags, lines with low amount of alphabetic symbols etc. We also remove lines which appear to be emails or web-addresses and duplicates.

- Language identification
  The algorithm is a fairly simple ensemble of three tools: `pycld2`[1], `langid` (Lui and Baldwin, 2012), `langdetect`[2]. We mainly use pycld2 as it is by far the fastest tool of the three. If pycld2's output differs from the hint language, we perform additional checks using the other two libraries, and the final language is determined by majority vote. For large monolingual corpora we use only pycld2.

- Bicleaner filtering
  We use the bicleaner (Ramírez-Sánchez et al., 2020) tool to filter parallel data. We discard all sentence pairs with the score threshold <= 0.3.

- Scoring with NMT models
  We finally score all parallel data and back-translations with our intermediate models. We use a score threshold to discard a portion of the data. The exact threshold is determined by human evaluation. The discarded data includes non-parallel sentences (i.e. pairs of sentences where the source does not correspond to the target in part or fully) and low-quality synthetic translations.

- Dual conditional cross-entropy filtering
  This year we use this algorithm again for both directions as described in Junczys-Dowmunt (2018).

2.2 English-Russian

The English-Russian system is basically the same as last year (Molchanov et al., 2022). It was trained in two steps. First, we build the baseline model on all available data. Second, we fine-tune the model on data of high quality. Specifically, we remove the ParaCrawl, UN and OpenSubtitles corpora. The training corpus then consists of the remains of the human data mixed with the back-translations of the news corpora (2020, 2021) from statmt.org. This approach shows good results according to automatic metrics and general translation quality. The reason for doing this is that we aim for our models to be used mostly for translation of news and formal texts like various types of documents.
The system was trained with separate vocabularies, the sizes of the BPE models are 24k for the source side and 48k for the target side.

[1] https://pypi.org/project/pycld2/
[2] https://pypi.org/project/langdetect/

source	Model2022	Model2023	Model2023 fixed
Перенасадка башмаков и колец для колпаков, замена вентилей должны	Overpressure of shoes and rings for caps, replacement of valves shall	\\\\\\\{ 1\\\\\\\> ·\\\\\\\{ 1\\\\\\\}\\\\\\\{ 2\\\\\\\}\\\\\\\< 3\\\\\\\}	Re-fitting of shoes and rings for caps, replacement of valves should
Прогноз компонентов ВВП по использованию на 2019 г. несколько изменился, что связано прежде всего с выходом фактических данных за II квартал 2019 года.	The forecast of GDP components for use for 2019 has changed somewhat, which is primarily due to the release of actual data for the second quarter of 2019.	2019 GDP Components Forecast by Usage The Bank of Russia's monetary policy is based on the following principles:	The forecast of GDP components for 2019 slightly changed, which is primarily due to the release of actual data for the second quarter of 2019.

Table 2: Examples of degradations for the 2023 Russian-English model.

2.3 Russian-English

The Russian-English model was built basically on the same data and in the same way as the English-Russian model. The only difference is that we use English news and Wikipedia for back-translations. The previous version of the Russian-English model was also built on the same data, but with the transformer-base configuration.

The first version of the model that we trained on this data had shown almost no improvements, both in terms of automatic and human evaluation (on average the model improved by 0.5 BLEU points on our internal test sets compared to Model2022 using the transformer-base configuration). What is more important is that we observed some serious degradations: hallucinations and critical mistakes. The examples are presented in Table 2. We investigated the problem and found out that some of our clients' data was used for training without proper filtering. This was part of our private data. We then applied the full filtering pipeline to the private data and discarded around 20k sentence pairs (roughly 0.03% of all data) with low quality. Then we retrained the model on the filtered data, and this fixed all the critical mistakes we had encountered. Surprisingly, we also gained additional 1 BLEU points on average on our internal test sets compared to the first version. All we did was just remove 0,03% of bad sentence pairs from the training data. The average BLEU score on our test sets improved from 36.66 to 38.05 points.

3 Results and discussion

The results are presented in Table 3.

As we can see, we outperform our baselines (i.e. previous versions of the models). The gains we observe, however, are not that large.

System	BLEU	chrF	COMET
English-Russian			
Model2022	30.5	55.4	82.3
Russian-English			
Model2022	32.4	58.0	-
Model2023	**32.8**	**58.4**	**80,9**

Table 3: Results for different systems in both directions. The submitted systems are marked in bold. Model2022 stands for our previous version of the Russian-English system which we consider the baseline. The English-Russian system remains the same.

However, other test sets, such as the TICO-19 evaluation set [3] (Anastasopoulos et al., 2020), show more substantial improvements. The BLEU score on that test set has grown from 33.8 to 35 points.

Poor performance on the generaltest2023 set can be due to the problems that our submitted models have with translation of colloquial content. This can be explained by our data preparation scheme. As we have already mentioned above, we want our models to translate formal text better and thus 'sacrifice' colloquial data. The examples of such mistranslations are presented in Table 4. Both examples illustrate colloquial slang which our model cannot translate properly. In the first example the word 'please' is substituted by 'pls', and thus the model 'thinks' it is a abbreviation of some kind. In the second example the author substitutes the word 'because' with a slang word 'becuz', and the model transliterates it.

We made a thorough investigation into the generaltest2023 sets. Thus, we found out that there are four major topics for the Russian-English test set: 1) movie reviews; 2) news of any

[3] https://tico-19.github.io/index.html

kind; 3) user reviews; 4) abstracts from research papers in medical domain. The English-Russian test set domains are similar. We estimate that at least half of the English-Russian test set is made up of Reddit posts and online customer reviews which often use internet slang and have spelling anomalies of some kind, e.g. "eye wud liek 2 aply 4 vilage idot", "WhY dO pPl FiNd ThE NeW SeT ExPeNsIvE." All these domains except for news were actually unexpected by our model.

source	Model2022
pls change	Изменение PLS
Becuz i have less than 3000 points.	Бекуз у меня меньше 3000 очков.

Table 4: Examples of incorrect translations for the English-Russian model considering the colloquial content.

4 Conclusions and future work

In this paper we presented our submissions for the WMT23 Shared General Translation Task. We show good results in both directions we participate. We clearly outperform our baselines in both directions. A detailed analysis of the translations shows us that we lose quality in translation of colloquial speech. We have already started to work in this direction. We have synthesized data where, e.g., 'please' is substituted with 'plz' and so on. We plan to train our model on this synthetic data so that it could deal with such colloquial examples.

References

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged Back-Translation. In *Proceedings of the Fourth Conference on Machine Translation*, pages 53–63, Florence, Italy.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of NAACL HLT 2013*, pages 644–648, Atlanta, USA.

Marcin Junczys-Dowmunt. 2018. Dual Conditional Cross-Entropy Filtering of Noisy Parallel Corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast Neural Machine Translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, Alexander M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *Computing Research Repository*, arXiv:1701.02810. Version 2.

Marco Lui and Timothy Baldwin. 2012. langid.py: An Off-the-shelf Language Identification Tool. In *Proceedings of ACL 2012, System Demonstrations*, pages 25–30, Jeju, Republic of Korea.

Alexander Molchanov. 2019. PROMT Systems for WMT 2019 Shared Translation Task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 302–307, Florence, Italy.

Alexander Molchanov, Vladislav Kovalenko, Natalia Makhamalkina. 2022. PROMT Systems for WMT22 General Translation Task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 342–345, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and WeiJing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 02)*, pages 311–318, Philadelphia, PA, USA.

Marcis Pinnis, Rihards Krišlauks, Daiga Deksne, and Toms Miks. 2017. Neural Machine Translation for Morphologically Rich Languages with Improved Sub-word Units and Synthetic Data. In *Proceedings of the 20th International Conference of Text, Speech and Dialogue (TSD2017)*, pages 237–245, Prague, Czechia.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón and Sergio Ortiz Rojas. 2020. Bifixer and Bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, Berlin, Germany.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with

subword units. 2016b. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 35–40, Berlin, Germany.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the Translation Initiative for COvid-19. In Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020, Online. Association for Computational Linguistics.

AIST AIRC Submissions to the WMT23 Shared Task

Matīss Rikters[1]
[1]Artificial Intelligence Research Center (AIRC)
National Institute of Advanced Industrial Science and Technology
matiss.rikters@aist.go.jp

Makoto Miwa[1,2]
[2]Toyota Technological Institute, Japan
makoto-miwa@toyota-ti.ac.jp

Abstract

This paper describes the development process of NMT systems that were submitted to the WMT 2023 General Translation task by the team of AIST AIRC. We trained constrained track models for translation between English, German, and Japanese. Before training the final models, we first filtered the parallel and monolingual data, then performed iterative back-translation as well as parallel data distillation to be used for non-autoregressive model training. We experimented with training Transformer models, Mega models, and custom non-autoregressive sequence-to-sequence models with encoder and decoder weights initialised by a multilingual BERT base. Our primary submissions contain translations from ensembles of two Mega model checkpoints and our contrastive submissions are generated by our non-autoregressive models.

1 Introduction

We describe the machine translation (MT) systems submitted to the WMT 2023 General Translation task developed by the team of AIST AIRC. We experimented with data quality control by carefully filtering out noisy examples from parallel and monolingual data sets before training, and corpora selection by holding out specific web-crawled data. We also compared several modelling approaches by contrasting the well-known Transformer architecture (Vaswani et al., 2017) to several more recent ones, such as the Mega model (Ma et al., 2023), as well as our own custom implementation of a non-autoregressive model with the encoder and decoder initialised by BERT checkpoints. During the shared task submission week another new efficient architecture was published – the Retentive Network (RetNet; Sun et al., 2023), which we include in the paper as an ablation study.

Our main findings are: 1) non-autoregressive models can reach comparable output quality to the best autoregressive models while improving inference latency up to 9x; 2) modern efficient autoregressive models like RetNet and Mega not only slightly outperform the Transformer in latency, but also in output quality; and 3) models trained on sentence-level data struggle to translate whole paragraphs – splitting them into sentences helps a lot, especially for the non-autoregressive model.

2 Data

We only participated in the constrained track of the shared task; therefore, we limited our data set use to only the corpora provided by the shared task organisers. In specific experimentation configurations, we chose to leave out web-crawled data such as Paracrawl and WikiMatrix, but eventually kept them in our final submissions.

All parallel training data and monolingual data for back-translation were filtered before starting any training, which has been proven very effective in previous WMT shared tasks (Pinnis et al., 2017, 2018) and detailed by Rikters (2018). Parallel data distillation was performed only for training the non-autoregressive models, while for all autoregressive models, we used only pure clean parallel data.

For the system development process, we selected News Test sets from previous older WMT shared tasks as development data and the most recent ones as evaluation data. Full statistics of the data we used are shown in Table 1.

2.1 Data Selection

We initially experimented with excluding the web-crawled parallel corpora and training models using only data from other sources, since web-crawled data are generally considered to be of a lower-quality tier. The Paracrawl corpora are also several times the size of all other data combined, and took longer to finish the filtering process. In addition, to not overwhelm the full combined training data set with lower-quality data, we 1) limited the English-

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 155–161
December 6–7, 2023.

Corpus / Filtering		DE-EN	JA-EN
All other	Before	16,752,302	8,076,155
	After	13,737,028	7,076,869
Paracrawl	Before	50,000,000	21,891,738
	After	44,533,635	21,088,689
	Combined	72,007,691	42,319,296

Devel	19,006	2,998
Eval	3,039	3,037

	Monolingual	
Corpus / Filtering	Before	After
DE	43,613,631	37,110,981
JA	22,193,545	21,558,123
EN	47,333,840	36,756,542

Table 1: Training data statistics for all other parallel data without Paracrawl, a subset of Paracrawl, combined development and evaluation data from the past WMT shared tasks, and monolingual data. Sentence counts are listed before and after filtering.

German Paracrawl to 50 million parallel sentences; and 2) up-scaled all data from other sources to match the amount of the Paracrawl data after filtering by doubling for English-German and tripling for English-Japanese.

2.2 Filtering

Even though all training data need not always be perfect and methods like back-translation and data distillation intentionally generate somewhat noisy additional training data, some types of noise are more harmful than others. Since most training corpora are produced partially or fully automatically, errors such as misalignments between source and target sentences or direct copies of source to target can occur, as well as some amounts of third language data in seemingly bilingual data sets.

To avoid such problems, we used data cleaning and pre-processing methods described by Rikters (2018). The filtering part includes the following filters: 1) unique parallel sentence filter; 2) equal source-target filter; 3) multiple sources - one target and multiple targets - one source filters; 4) non-alphabetical filters; 5) repeating token filter; and 6) correct language filter. We also perform pre-processing consisting of the standard Moses (Koehn et al., 2007) scripts for punctuation normalisation, cleaning, and Sentencepiece (Kudo and Richardson, 2018) for splitting into subword units. The filters were applied to the given parallel sentences, monolingual news sentences before performing back-translation, and both sets of synthetic parallel sentences resulted from back-translating the monolingual news.

2.3 Distillation

Since previous research has proven that knowledge distillation (Hinton et al., 2014) is highly beneficial for non-autoregressive machine (NAR) translation models (Kim and Rush, 2016), we chose to skip training our NAR models during the baseline training phase. When the baselines were trained, evaluated and compared, we used the highest-scoring baseline models for sentence-level knowledge distillation of the clean parallel training data.

2.4 Back-translation

Increasing the amount of in-domain training data with synthetic back-translated corpora (Sennrich et al., 2016) has become a common practice in cases with considerable amounts of in-domain monolingual data. However, since the shared task recently shifted from 'news' to 'general' text translation, the definition of what would be considered in-domain data became less clear. Furthermore, for the constrained track the selection of provided monolingual data from the organisers was limited to news and web-crawled data while noting that the 'general' test sets may include user generated (social network), conversational, and e-commerce data as well. For our experiments we continued to assume that a significant portion of the test data would still be from the news domain. Therefore, we chose to only use the provided monolingual News crawl, News discussions, and News Commentary corpora for back-translation.

2.5 Post-processing

In post-processing of the model output we aimed to mitigate some of the most commonly noticable mistakes that the models were generating. We mainly noticed two often occurring problems in output from all models: 1) difficulties in translating emoji symbols; and 2) occasional repetitions of words or phrases.

While all English and German alphabet letters and even Japanese characters are covered in the large training data corpora, the unicode emoji were mostly formed and clearly defined only in the past decade, and new emoji are still added every year or two with the next release planned for late

2024[1]. Emoji are also not often present in MT training data, therefore full emoji coverage is absent from model vocabularies, which leads to occasional <*unk*> tokens being generated as output if emoji were present in the input. In order to keep using the models without re-training, we replaced any <*unk*> tokens in the output using a dictionary of any emojis appearing in the input.

Furthermore, the occasional hickuping or hallucinating of models on less common input sequences seems ever present, sometimes generating repetitions of tokens or phrases. We replaced any consecutive repeating n-grams with a single n-gram. The same was applied to repeating n-grams that have a preposition between them, i.e., *the victim of the victim*.

Both post-processing approaches gave BLEU score improvements of around 0.1 - 0.2.

3 Model Configurations

While it is often possible to train ever larger models on more data requiring infinitely growing amounts of compute power which later become costly to deploy, we decided to approach our selection from the perspective of limiting environmental impact. In our pursuit of the final submission, we aimed to explore several modelling approaches with efficient decoding while still striving to maintain or improve output quality. For this we chose the baseline Transformer model as our baseline, the recently introduced Mega model (Ma et al., 2023), a custom implementation of a non-autoregressive model with BERT-initialised encoder and decoder, and as an ablation study trained after the shared task submission deadline – RetNet (Sun et al., 2023). Each model was trained on a single machine with four Nvidia V100 (16GB) GPUs until convergence on development data (no improvement on validation loss for 7 checkpoints).

The total trainable parameter counts for the four models are as follows: Transformer - 73,886,208; RetNet - 77,930,496; Mega - 63,367,854; BnB - 384,214,027.

3.1 Transformer

We used Marian (Junczys-Dowmunt et al., 2018) to train transformer architecture (Vaswani et al., 2017) models with the default transformer-base parameter configuration of 6 layers, 8 attention heads, model dimension of 512, feed-forward dimension of 2048, and dropout of 0.1. We also used an optimiser delay of 8 to simulate larger batches, which is is known to improve final output quality (Bogoychev et al., 2018).

3.2 Mega

Ma et al. (2023) propose a moving average equipped gated attention mechanism (MEGA) - a single-head gated attention mechanism equipped with exponential moving average to incorporate inductive bias of position-aware local dependencies into the position-agnostic attention mechanism. Compared to the Transformer model, MEGA has a single-head gated attention mechanism instead of multi-head attention, which enables gains in efficiency while not sacrificing on performance.

For training our Mega models we used the implementation[2] provided by the authors, which is based on FairSeq (Ott et al., 2019).

3.3 BERT-nar-BERT

The BERT-nar-BERT (BnB) model architecture is similar to BioNART (Asada and Miwa, 2023), composed of a multi-layer Transformer-based encoder and decoder, in which the embedding layer and the stack of transformer layers are initialised with BERT (Devlin et al., 2019). To leverage the expressiveness power of existing pre-trained BERT models, we initialise our encoder and decoder parts with the pre-trained BERT parameters. An overview of BnB architecture is shown in Figure 1.

The encoder part of BnB is the same architecture as the BERT model. We construct latent representations based on token-level representations from the encoder hidden state, and modify the decoder part by leveraging the latent representations and length classification for non-autoregressive generation.

The decoder part is also based on the BERT architecture, and we can directly initialise the decoder with the pre-trained BERT model. Following the BERT2BERT model, the cross-attention mechanism is adopted, and the encoder hidden representation of the final layer is used for cross-attention. Our model differs from the BERT2BERT model in attention masks to enable NAR decoding. In the AR decoding, all target tokens are fed into the decoder with customised attention masks that prevent the decoder from seeing the future tokens during training. Then, in inference, the predicted token is fed to the decoder autoregressively. In our BnB de-

[1] https://emojipedia.org/unicode-16.0

[2] https://github.com/facebookresearch/mega

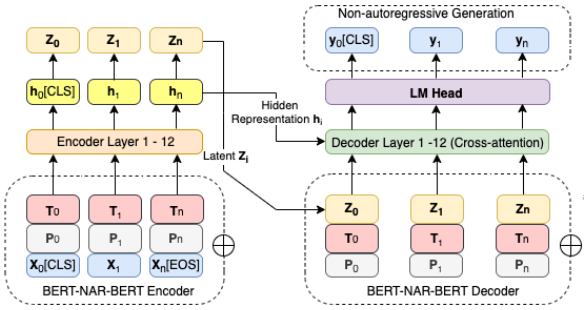


Figure 1: The S2S BERT-nar-BERT (BnB) architecture.

Model	GPU	CPU	Speedup
Transformer	30.08	4.71	1.00x
MEGA	43.67	6.81	1.45x
RetNet	43.42	6.99	1.46x
BnB	278.83	13.23	6.04x

Table 2: Average speedup and inference speed in lines per second on CPU and GPU on average for the four WMT 2023 test sets we participated in.

coder, input representation is constructed without providing any target tokens. The input representation is constructed by summing the corresponding position and type embeddings and the latent embedding from the encoder. The attention masks are the normal masks that give access to all future tokens. The resulting decoder output representations of the final layer are fed to the subsequent generation layer.

3.4 Ablation Study – Retentive Networks

During the submission week of the WMT general machine translation task Sun et al. (2023) proposed a Retentive network (RetNet), with stacked identical blocks, following a similar layout to the Transformer, where each block contains a multi-scale retention module, and a feed-forward network module. Compared to Transformer attention, the retention part removes softmax and enables recurrent formulation, which significantly benefits inference. The authors report significant gains in inference efficiency while maintaining competitive in output quality to the Transformer.

For training our RetNet models we used the implementation[3] provided by the authors, which is based on FairSeq (Ott et al., 2019).

4 Results

Tables 3 and 4 list the progression of our different modelling methods and data selection approaches. We first started with training the Transformer models as our baselines using only non-web-crawled parallel training data and compared it to MEGA models trained on the same data, while the larger Paracrawl corpora were still filtering. Initial results suggested that the Transformer model optimises towards the development data slightly too much while ending up strongly outperformed by the MEGA model on evaluation data. From there on, we opted for using MEGA as our main model, and experimented with adding filtered Paracrawl data to the training mix, which improved translation quality for all directions. We then used these four models (With Paracrawl column in Table 3) to generate back-translated data and distilled parallel training data for BnB. In the final step before submission, we trained MEGA and BnB models on clean parallel + back-translated and distilled + back-translated data respectively. We used ensembles of best and last MEGA model checkpoints to generate our shared task submissions.

As an ablation study of adding another efficient model baseline, after the submission week had ended we trained RetNet models, which were published on arXiv along with code on GitHub during the submission week.

4.1 Automatic Evaluation

According to the unofficial automatic evaluation results (Kocmi et al., 2023) summarised in Table 6, our submitted models are on the lower end, outperforming only two to three out of the 5-10 participants and 7 online systems in the respective translation directions. We manually regenerated the automatic evaluation scores for translations from all of our final models, based on the references released by the organisers.

4.2 Inference Speed

Table 2 compares the inference speed and latency of our chosen models. While loading the models into the memory and model-specific data preprocessing or post-processing steps also take considerable amounts of time, for this comparison we only started measuring the time after the model had been loaded and all data processing – completed. Our BnB model was by far the fastest, outperforming MEGA and Retnet by about 6.4x on the GPU and the Transformer by about 9.3x. On the CPU

[3] https://github.com/microsoft/torchscale

Direction	Without Paracrawl				With Paracrawl		Back-translated	
	Transformer		MEGA					
	Devel	Eval	Devel	Eval	Devel	Eval	Devel	Eval
EN→DE	**32.74**	19.46	28.96	25.15	31.42	28.04	31.58	**26.91**
DE→EN	34.57	22.13	30.55	26.22	34.67	29.21	**36.62**	**27.85**
EN→JA	20.01	7.13	16.52	16.07	19.29	21.00	**20.89**	**20.90**
JA→EN	15.42	5.98	13.39	12.27	16.82	16.15	**17.43**	**16.12**

Table 3: Initial baseline Transformer and Mega model development results using filtered parallel data excluding Paracrawl, all filtered parallel data, and all filtered parallel data + back-translated data.

Direction	MEGA Ensembles				RetNet				BnB	
	Back-translated		All Filtered		Ensemble BT		Back-translated		Distilled + BT	
	Devel	Eval	Devel	Eval	Devel	Eval	Devel	Eval	Devel	Eval
EN→DE	32.33	27.52	**32.51**	**28.76**	31.92	27.10	31.99	27.25	25.34	22.40
DE→EN	**37.56**	28.50	35.35	**29.62**	37.44	28.49	37.17	28.14	28.04	24.23
EN→JA	**21.31**	21.13	18.98	**21.23**	21.67	21.87	21.64	21.64	11.45	13.38
JA→EN	**18.08**	**16.81**	17.19	16.23	18.10	17.10	18.36	17.26	7.93	8.03

Table 4: MEGA, our BnB model, and RetNet model development results using all filtered data, back-translated data, and ensembles of trained model checkpoints. A combination of back-translated monolingual data and distilled parallel data was used to train our BnB model. Highest scores reached before the shared task submission deadline are marked in bold and after the deadline – underlined.

Direction	MEGA	BnB	RetNet	Transformer
EN→DE	26.48	5.58	29.31	26.11
Split	34.30	29.93	34.89	35.57
DE→EN	32.35	15.98	34.04	32.02
Split	37.14	30.10	37.57	39.52
EN→JA	17.28	15.25	17.44	14.76
JA→EN	18.53	6.96	15.34	17.64

Table 5: Final results on *GeneralTest2023* after the shared task submission deadline.

its advantage dropped to about 1.9x and 2.8x respectively. Inference speed differences between MEGA and RetNet were minimal, while both still noticabely outperformed the baseline Transformer.

4.3 Post Submission Updates

After the release of the unofficial system rankings and test set references, we manually re-scored all of our models trained on the final back-translated data and noticed that the Transformer and BnB were generating particularly shorter outputs for the document-level EN↔DE test sets than expected. After splitting[4] the English and German source files into sentences, translating them, and combining back into paragraphs for evaluation, the scores improved by several BLEU points (see Table 5). The EN↔JA part did not require any further splitting, as it was already provided at sentence-level.

5 Conclusion

In this paper we described the development process of the AIST AIRC's NMT systems that were submitted for the WMT 2023 shared task on general domain text translation. We compared Transformer models to MEGA, RetNet and BERT-nar-BERT model architectures in search of efficient decoding approaches while still improving upon output quality. We showed that the Transformer models can be outperformed by MEGA and RetNet in both translation quality, as well as inference speed, while BnB remained fastest in inference, but still lowest in quality. We also found that even though modern models should be able to handle long sequences, splitting the English↔German document-level data into separate sentences, translating and recombining them yielded better results. This should, however, be mitigable by training dedicated document-level models with appropriate training data.

In total, output from four systems was submitted to the shared taks by AIRC for the English↔German and English↔Japanese language pairs in both translation directions.

[4]Text to Sentence Splitter – https://github.com/mediacloud/sentence-splitter

System	BLEU
ONLINE-W	51.8
GPT4-5shot	47.9
ONLINE-A	47.9
ONLINE-B	46.3
ONLINE-G	46.0
ONLINE-Y	43.9
GTCOM_Peter	42.2
Lan-BridgeMT	42.1
ONLINE-M	41.3
ZengHuiMT	40.8
NLLB_Greedy	33.1
AIRC	32.4
NLLB_MBR_BLEU	32.4

System	BLEU
ONLINE-W	47.8
ONLINE-A	43.7
GPT4-5shot	43.6
ONLINE-Y	43.6
ONLINE-G	43.2
ONLINE-B	42.7
ONLINE-M	40.5
ZengHuiMT	40.5
Lan-BridgeMT	39.4
NLLB_Greedy	31.1
NLLB_MBR_BLEU	29.6
AIRC	26.5

System	BLEU
ONLINE-W	25.9
SKIM	24.8
GPT4-5shot	24.1
ONLINE-B	23.9
NAIST-NICT	23.0
ONLINE-A	23.0
ZengHuiMT	22.6
GTCOM_Peter	22.3
ONLINE-Y	22.3
ANVITA	20.9
Lan-BridgeMT	20.2
ONLINE-G	18.3
KYB	17.6
ONLINE-M	17.2
AIRC	14.9
NLLB_MBR_BLEU	14.7
NLLB_Greedy	14.2

System	BLEU
ONLINE-B	25.3
ONLINE-W	24.5
ONLINE-Y	24.5
SKIM	24.3
NAIST-NICT	22.6
ZengHuiMT	22.6
ONLINE-A	21.4
GPT4-5shot	21.3
Lan-BridgeMT	20.5
ONLINE-M	19.8
ANVITA	19.4
KYB	17.8
AIRC	17.6
ONLINE-G	17.2
NLLB_Greedy	11.3
NLLB_MBR_BLEU	9.0

System	Chr F
ONLINE-W	72.1
ONLINE-A	70.0
GPT4-5shot	69.8
ONLINE-B	69.1
ONLINE-G	69.1
ONLINE-Y	68.4
ZengHuiMT	67.6
Lan-BridgeMT	66.7
GTCOM_Peter	66.6
ONLINE-M	66.5
NLLB_MBR_BLEU	57.6
NLLB_Greedy	57.3
AIRC	57.2

System	Chr F
ONLINE-W	71.8
ONLINE-A	69.7
ZengHuiMT	69.4
GPT4-5shot	69.1
ONLINE-B	69.1
ONLINE-Y	69.1
ONLINE-G	69.0
ONLINE-M	66.9
Lan-BridgeMT	66.1
NLLB_Greedy	56.2
NLLB_MBR_BLEU	55.4
AIRC	52.2

System	Chr F
ONLINE-W	51.4
GPT4-5shot	51.2
SKIM	51.1
ONLINE-A	49.6
NAIST-NICT	49.5
ONLINE-Y	49.5
ZengHuiMT	49.5
ONLINE-B	49.3
GTCOM_Peter	48.7
Lan-BridgeMT	47.3
ANVITA	46.7
ONLINE-G	45.5
KYB	43.9
ONLINE-M	43.9
AIRC	40.5
NLLB_MBR_BLEU	39.2
NLLB_Greedy	39.0

System	Chr F
ONLINE-B	35.2
ONLINE-Y	34.1
ONLINE-W	33.5
SKIM	33.5
ZengHuiMT	32.9
NAIST-NICT	32.0
ONLINE-A	31.4
GPT4-5shot	31.0
Lan-BridgeMT	30.4
ONLINE-M	29.6
ANVITA	29.3
KYB	27.7
AIRC	27.6
ONLINE-G	27.3
NLLB_Greedy	20.9
NLLB_MBR_BLEU	18.7

System	COMET
GPT4-5shot	86.3
ONLINE-W	86.0
ONLINE-B	85.6
ONLINE-A	85.5
ONLINE-Y	84.9
ONLINE-M	84.8
ONLINE-G	84.6
GTCOM_Peter	82.7
NLLB_MBR_BLEU	81.4
ZengHuiMT	81.1
Lan-BridgeMT	80.9
NLLB_Greedy	79.9
AIRC	78.7

System	COMET
ONLINE-W	85.5
GPT4-5shot	85.0
ONLINE-B	84.8
ONLINE-Y	84.1
ONLINE-A	83.7
ONLINE-G	82.5
ONLINE-M	81.7
Lan-BridgeMT	80.4
ZengHuiMT	79.4
NLLB_MBR_BLEU	78.0
NLLB_Greedy	77.9
AIRC	72.9

System	COMET
SKIM	84.0
GPT4-5shot	83.4
ONLINE-W	82.3
NAIST-NICT	81.9
ONLINE-Y	81.6
ONLINE-B	81.5
ONLINE-A	81.0
GTCOM_Peter	80.2
ANVITA	79.5
Lan-BridgeMT	79.3
ZengHuiMT	79.2
ONLINE-G	77.8
ONLINE-M	77.5
KYB	76.6
NLLB_MBR_BLEU	75.2
AIRC	74.5
NLLB_Greedy	74.3

System	COMET
ONLINE-B	88.2
ONLINE-W	87.5
ONLINE-Y	87.3
GPT4-5shot	87.0
SKIM	86.6
NAIST-NICT	86.2
ZengHuiMT	85.3
ONLINE-A	85.2
Lan-BridgeMT	84.5
ONLINE-M	13.3
ANVITA	82.7
KYB	80.8
AIRC	80.7
ONLINE-G	80.4
NLLB_Greedy	79.3
NLLB_MBR_BLEU	77.7

Table 6: Automatic evaluation rankings according to BLEU (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.2.1), chrF (nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.2.1), and COMET (Unbabel/wmt22-comet-da). The order of the tables from left to right is DE→EN, EN→DE, JA→EN, EN→JA.

In future work, we plan to experiment with replacing the BERT models in BnB with other more efficient pre-trained language models which can be used as encoders/decoders, as well as incorporating document-level training data and modelling longer sequences with available data. In terms of data, we intend to increase vocabulary coverage by adding all known unicode emoji symbols to the vocabulary even if they are not present in the training data, as well as additionally sample paracrawl data where emoji are present.

Acknowledgements

This paper is based on results obtained from a project JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

Ethics Statement

Our work fully complies with the ACL Code of Ethics[5]. We use only publicly available datasets and relatively low compute amounts while conducting our experiments to enable reproducibility. We do not perform any studies on other humans or animals in this research.

References

Masaki Asada and Makoto Miwa. 2023. BioNART: A biomedical non-AutoRegressive transformer for

[5] https://www.aclweb.org/portal/content/acl-code-ethics

natural language generation. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 369–376, Toronto, Canada. Association for Computational Linguistics.

Nikolay Bogoychev, Kenneth Heafield, Alham Fikri Aji, and Marcin Junczys-Dowmunt. 2018. Accelerating asynchronous stochastic gradient descent for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2991–2996, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2014. Distilling the knowledge in a neural network. *NIPS 2014 Deep Learning Workshop*.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Xuezhe Ma, Chunting Zhou, Xiang Kong, Junxian He, Liangke Gui, Graham Neubig, Jonathan May, and Luke Zettlemoyer. 2023. Mega: Moving average equipped gated attention. In *The Eleventh International Conference on Learning Representations*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Mārcis Pinnis, Rihards Krišlauks, Toms Miks, Daiga Deksne, and Valters Šics. 2017. Tilde's machine translation systems for WMT 2017. In *Proceedings of the Second Conference on Machine Translation*, pages 374–381, Copenhagen, Denmark. Association for Computational Linguistics.

Mārcis Pinnis, Matīss Rikters, and Rihards Krišlauks. 2018. Tilde's machine translation systems for WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 473–481, Belgium, Brussels. Association for Computational Linguistics.

Matīss Rikters. 2018. Impact of Corpora Quality on Neural Machine Translation. In *In Proceedings of the 8th Conference Human Language Technologies - The Baltic Perspective (Baltic HLT 2018)*, Tartu, Estonia.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. 2023. Retentive network: A successor to transformer for large language models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

MUNI-NLP Submission for Czech-Ukrainian Translation Task at WMT23

Pavel Rychlý and **Yuliia Teslia**
NLP Centre, Faculty of Informatics, Masaryk University
pary@fi.muni.cz, 531354@mail.muni.cz

Abstract

The system is trained on officially provided data only. We have heavily filtered all the data to remove machine translated text, Russian text and other noise. We use the DeepNorm modification of the transformer architecture in the TorchScale library with 18 encoder layers and 6 decoder layers. The initial systems for back-translation uses HFT tokenizer, the final system uses custom tokenizer derived from HFT.

1 Introduction

The annual Conference in Machine Translation (WMT) provides an invaluable platform for researchers to showcase their advancements in this domain. This paper serves as the submission from the Natural Language Processing Centre of Masaryk University (MUNI-NLP) team for the Czech-Ukrainian Translation Task at WMT23 (Kocmi et al., 2023).

Central to our approach is a commitment to data quality and we have also done some experiments with different subword tokenizers. Furthermore, we recognize the paramount importance of data filtering, a process that distinguishes signal from noise. To this end, we employ a rigorous data filtering procedure to eliminate machine-translated text, Russian content, and other sources of potential distraction, enabling our model to focus on the genuine linguistic intricacies of Czech and Ukrainian.

In terms of model architecture, we employ the DeepNorm modification (Wang et al., 2022) of the transformer architecture. This modified architecture, integrated within the TorchScale library, boasts 18 encoder layers and 6 decoder layers.

Our approach further delves into the critical aspects of back-translation and tokenization. Initially, we leverage the HFT (High Frequency Tokens) tokenizer for back-translation, harnessing the power of synthetic data to enhance model robustness. In the final iteration of our system, we introduce a custom tokenizer derived from HFT.

2 Data selection and preprocessing

Our system participates in the constrained track, we use only data allowed for this year. We do not use any pretrained models, only selected parallel and monolingual texts.

Many of the provided text are very noisy. We excluded some of them from training completely.

To mitigate the adverse effects of noisy data on our translation system, we conduct a comprehensive analysis of the data problems that emerge from these issues. In subsequent sections of this paper, we delve into the specific strategies and techniques we employ to filter out machine-translated text, Russian content, and other sources of noise. Our data filtering pipeline is designed to rigorously curate the training data set, ensuring that our model is exposed to high-quality, human-generated translations that align closely with the nuances of the Czech and Ukrainian languages. Through these meticulous filtering strategies, we aim to enhance the overall performance and translation quality of our system.

2.1 Parallel data

We use only official Task 1 data downloaded by the `mtdata` command (Gowda et al., 2021). The majority of segments comes from the OPUS corpus (Tiedemann, 2012), the biggest single source is Facebook-wikimatrix (Schwenk et al., 2019).

OPUS-opensubtitles Sometimes contains wrong or missing diacritics in Czech part. We removed segments containing meta data like authors of the subtitles. There are many parts on the Ukrainian side with Russian language instead of Ukrainian. We have removed such segments.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 162–165
December 6–7, 2023.

Source	Used segments	Original segments	Used words
ELRC-acts-ukrainian	130003	130003	2.5M
OPUS-ccmatrix	3916740	3991954	44M
OPUS-opensubtitles	515216	730804	2.7M
OPUS-multiparacrawl	941349	2200276	12M
OPUS-qed	155346	161020	2M
OPUS-tatoeba	2932	2933	11k
OPUS-ted2020	112689	114229	1.6M
Facebook-wikimatrix	824606	848961	9.9M
Total	6602828		

Table 1: The sizes of all sources used for the final system. *Used words* means number of words used in one language, these are almost same for both languages.

In total, almost 30% of segments were removed from this source.

Facebook-wikimatrix Many segments are not aligned, they contains similar texts but the sentences are not translations. We can see such situations in sentences about different sport teams, towns and history persons.

We have removed segments with special formatting options, lines containing *Dostupné online (available online)* and similar strings.

OPUS-wikimedia Removed HTML formatting, notes in tested parenthesis which are not translated anyway.

Removed segments containing URL, references to online sources.

Removed segments with Czech texts in Ukrainian part and vice versa.

OPUS-multiccaligned Excluded from processing.

The Czech part contains almost exclusively a very bad machine translations in domains like: game playing, health recommendations, porn, bitcoins, garden.

Only a few good Czech sentences are copied from Czech Wikipedia.

OPUS-bible Excluded from processing.

It contains very old language with unusual vocabulary and grammar.

OPUS-elrc-5179 Excluded from processing.

The same text as ELRC-acts with some errors (missing characters).

OPUS-eubookshop Excluded from processing.

Contains concatenated words on both sides.

Several sources contains duplicated segment, we keep only the first instance of such duplicates.

The sizes of all parallel sources used for the final system are listed in the Table 1.

2.2 Monolingual data

Statmt-news-crawl-2021-ces Removed time indications at the beginning of lines.

LangUk-* There is no punctuation in text.

We have used simple rules to add probable punctuation marks.

Removed markdown formatting.

We use additional filtering of back-translated segments. We use `filter-acktranslation.py` from (Popel et al., 2022). Unfortunately, there were still some Russian texts at the early stages of development and some Czech sentences were translated into Russian instead of Ukrainian. We filtered such segments out for the final system.

The sizes of monolingual data are listed in Table 2.

3 Tokenization

We use HFT tokenizer (Signoroni and Rychlý, 2022) for all stages. The tokenizer uses special characters to annotate word boundaries and character capitalization, they are listed in Table 3.

An example of tokenized text from the Czech part of the training data is in Figure 1. All uppercase letters are transformed to lower-case and the special characters preserve the original format. White spaces around punctuation marks are annotated explicitly in the same way as in Sentencepiece

```
A on: "Ne, tady jsme už asi rok."
↑¦a¦ ¦on¦ :▁" ↑¦ne¦ ,▁ ¦tady¦ ¦jsme¦ ¦už¦ ¦asi¦ ¦rok¦ ."

Trvalo 17 dní, než US senát zakázal
↑ ¦trvalo¦ ¦17¦ ¦dní¦ ,▁ ¦než¦ Δ¦us¦∇ ¦sen át¦ ¦zak ázal¦

Ona napsala: " MŮJ BANKÉŘ BUDE POTŘEBOVAT PŘEVOD 3 000$.
↑¦ona¦ ¦napsala¦ :▁ "▁ Δ ¦můj¦ ¦bank é ř¦ ¦bude¦ ¦potřebovat¦ ¦převo d¦∇ ¦3¦ ¦000¦ $ .
↑ ¦ona¦ ¦napsala¦ :▁ "▁ Δ ¦můj¦ ¦bank é ř¦ ¦bude¦ ¦potřebovat¦ ¦převo d¦ ∇ ¦3¦ ¦0 0 0¦ $ .
```

Figure 1: Example of the tokenization. The first line of each group is the plain text, the second line is the respective tokenization. The very last line is the modified tokenization used in the final system.

Source	Used segments
Leipzig-news	1M
Leipzig-newscrawl	1M
Leipzig-wikipedia	1M
Statmt-news-crawl.ces	11M
LangUk-ubercorpus	22M
LangUk-news	15M
Total UK	41M
Total CS	15M

Table 2: The sizes of all sources used for back-translation.

¦	<token-delimiter>
↑	<single-uppercase>
▁	<explicit-whitespace>
∇	<all-uppercase>
Δ	<end-of-uppercase>

Table 3: Special characters in the HFT tokenizations.

(Kudo and Richardson, 2018), but spaces between words are assumed as default.

For the final system we have made the following extra changes in tokenization:

- Separate special capitalization symbols from tokens, they are always separate tokens.
- Split numbers into digits.

An example of this changes is displayed on the very last line in the Figure 1, the first token is split into two tokens, the number "3 000" at the end of the line is tokenized into two tokens in the original tokenization and into four tokens in the final one.

For the initial systems for translation we use vocabulary size of 32,000 items. The final translations system use only 12,000 items in the vocabulary on each side.

These modifications in the final system were motivated by an experiment on smaller data where BLEU score increased from 22.4 to 24.9. Separating individual digits is also an option (disabled by default) in the Sentencepiece (Kudo and Richardson, 2018) tokenizer. We will do a detailed evaluation of this modifications in the future.

4 Model

We use the DeepNorm (Wang et al., 2022) modification of the transformer (Vaswani et al., 2017) architecture in the TorchScale library (Ma et al., 2022). Our early experiments with the number of encoder and decoder layers shows with the agreement of Wei et al. (2022) that asymmetric configuration with more encoder layer performs better. We use 18 encoder layers and 6 decoder layers in all our models.

The first stage of the system in CS-UK direction is trained only on parallel data (6.6M segments) for 30 epochs, second stage in UK-CS direction uses also 15M Czech segments and is trained for 17 epochs. The final system uses parallel data and back-translated Ukrainian monolingual data (41M segments). It is trained for only 4 epochs. Checkpoints are created every 2000 updates and the final submission is the average of 8 checkpoints (1 following and 6 preceding) around the top-scoring checkpoint on development data.

The performance of the individual models are detailed in Table 4.

Stage	direction	segments	BLEU
ST1	CS-UK	6.6M	31.17
ST2	UK-CS	19M	34.57
final	CS-UK	48M	35.87

Table 4: Progress of scores

5 Results

Our first submission evaluated by OCELoT system on the test data received very low scores (BLEU 15.6) which don't correlate to the scores on our development set. We noticed Russian sentences instead of Ukrainian in our translations. For the final submission, we have done more filtering of both parallel and monolingual (back-translated) data as described in Section 2. The same system on cleaned data received much better scores (BLEU 28.3)

The official scores of our final system on the test data (Kocmi et al., 2023) are listed in the Table 5.

	final	**first**
COMET	87.0	
chrF	57.0	41.0
BLEU	28.3	15.6

Table 5: Automatic Scores of the final system and the first submission.

6 Conclusion

This paper presents the MUNI-NLP submission to the WMT 2023 General Machine Translation Task. Our results show that it is very important to clean the training data, especially foreign languages.

The paper also introduces a novel tokenization into subwords, a detailed evaluation of it is part of our future work.

Acknowledgments

The work described herein has also been supported by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062 LINDAT/CLARIAH-CZ.

References

Thamme Gowda, Zhao Zhang, Chris Mattmann, and Jonathan May. 2021. Many-to-English machine translation tools, data, and pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 306–316, Online. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and Maja Popović. 2023. Findings of the 2023 conference on machine translation (WMT23). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Shuming Ma, Hongyu Wang, Shaohan Huang, Wenhui Wang, Zewen Chi, Li Dong, Alon Benhaim, Barun Patra, Vishrav Chaudhary, Xia Song, and Furu Wei. 2022. TorchScale: Transformers at scale. *CoRR*, abs/2211.13184.

Martin Popel, Jindrich Libovicky, and Jindrich Helcl. 2022. Cuni systems for the wmt 22 czech-ukrainian translation task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 352–357, Abu Dhabi. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*.

Edoardo Signoroni and Pavel Rychlý. 2022. HFT: High frequency tokens for low-resource NMT. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 56–63, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. 2022. DeepNet: Scaling Transformers to 1,000 layers. *CoRR*, abs/2203.00555.

Daimeng Wei, Zhiqiang Rao, Zhanglin Wu, Shaojun Li, Yuanchang Luo, Yuhao Xie, Xiaoyu Chen, Hengchao Shang, Zongyao Li, Zhengzhe Yu, Jinlong Yang, Miaomiao Ma, Lizhi Lei, Hao Yang, and Ying Qin. 2022. Hw-tsc's submissions to the wmt 2022 general machine translation shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 403–410, Abu Dhabi. Association for Computational Linguistics.

Exploring Prompt Engineering with GPT Language Models for Document-Level Machine Translation: Insights and Findings

Yangjian Wu
Lan-Bridge / Sichuan (China)
wuyangjian@lan-bridge.com

Gang Hu
Lan-Bridge / Sichuan (China)
hugang@lan-bridge.com

Abstract

This paper describes Lan-Bridge Translation systems for the WMT 2023 General Translation shared task. We participate in 2 directions: English to and from Chinese. With the emergence of large-scale models, various industries have undergone significant transformations, particularly in the realm of document-level machine translation. This has introduced a novel research paradigm that we have embraced in our participation in the WMT23 competition. Focusing on advancements in models such as GPT-3.5, we have undertaken numerous prompt-based experiments. Our objective is to achieve optimal human evaluation results for document-level machine translation, resulting in our submission of the final outcomes in the general track.

1 Introduction

Recently, large-scale language models, such as GPT-3.5 and GPT-4 (gpt), have emerged as powerful tools in the field of natural language processing. These models have showcased their impressive capabilities in a wide range of tasks, including text generation, question answering, language translation, and more. Language models like GPT-3.5 possess the ability to understand and generate coherent, contextually relevant text, capturing the nuances of language usage and producing high-quality outputs.

In particular, machine translation is an area where these language models have shown tremendous promise. Traditional machine translation models (Yang et al., 2020) used the conventional Transformer architecture (Vaswani et al., 2017) since GPT-3.5 has the potential to revolutionize the translation process by leveraging its massive size and language understanding capabilities. With the advent of large models, the machine translation field has faced new challenges, and utilizing large models for machine translation is a novel attempt.By effectively incorporating prompts and context, GPT-3.5 can produce translations that exhibit fluency, accuracy, and adherence to the source text.

This study focuses on experimenting and evaluating different prompt engineering techniques to further enhance the translation performance of GPT-3.5. By providing more refined and contextually specific prompts, we aim to observe the model's ability to adjust and refine its translations, resulting in improved translation quality. Additionally, we explore the impact of temperature adjustments on the generated translations, allowing us to fine-tune the level of randomness in the output and achieve more deterministic and accurate translations.

Furthermore, we investigate both sentence-level and document-level approaches, examining the effectiveness of GPT-3.5 in handling translations at different granularity levels. These approaches aim to leverage the model's language understanding capabilities to not only produce accurate sentence-level translations but also ensure coherence and consistency at the overall document level.

By delving into these aspects and evaluating the performance of GPT-3.5 in the context of the WMT competition, we aim to contribute to the broader understanding of the capabilities, strengths, and limitations of state-of-the-art language models in the field of machine translation.

The inspiration for this study stems from the outstanding performance exhibited by these large-scale language models, especially in addressing real-world challenges such as major wildfires. GPT-3.5-4k and GPT-3.5-16k, with their increased model capacities, have demonstrated remarkable capabilities in generating high-quality text across various domains. Motivated by these advancements, our study aims to harness the power of these models and explore their potential in the specific domain of machine translation.

By leveraging the robustness and adaptability of GPT-3.5-4k and GPT-3.5-16k, we conduct rigorous

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 166–169
December 6–7, 2023.

experimentation to thoroughly evaluate their translation capabilities. We delve into the nuances of different parameter adjustments, including prompts and temperature, to optimize and enhance the models' performance specifically for translation tasks. By strategically fine-tuning these parameters, we aim to unlock hidden potential and push the boundaries of their translation capabilities.

Real-world challenges, such as major wildfires, require timely and accurate translation of critical information across languages. The effectiveness of machine translation plays a pivotal role in communicating vital updates and ensuring efficient information dissemination during such situations. By investigating the translation capabilities of GPT-3.5-4k and GPT-3.5-16k, we strive to contribute insights that can improve translation efficiency and aid in overcoming language barriers in emergency situations.

With this study, we aim to shed light on the immense potential of large-scale language models, such as GPT-3.5, in addressing real-world challenges through machine translation. By harnessing their capabilities and understanding their performance in various scenarios, we hope to pave the way for more effective and accurate translation systems that can assist in critical situations.

2 Methods

We have designed three prompt schemes:

P1: Translate this sentence from SRC to TGT, do not write any explanations

P2: Translation Request - Sentence-by-Sentence Translation. Language Pair: SRC to TGT. Instructions: 1. Each sentence of the document will be provided individually in the "Original Sentence" section. 2. In the "Translation" section, please provide the corresponding translation for each sentence, considering the context and aiming for faithful translation while minimizing unaligned translations. 3. Avoid including any explanations in the translation. Original Sentence:

P3: Translation Request - Sentence-by-Sentence Translation. Language Pair: SRC to TGT. Instructions: 1. Each sentence of the document will be provided individually in the "Original Sentence" section. 2. In the "Translation" section, please provide the corresponding translation for each sentence, considering the context and aiming for faithful translation while minimizing unaligned translations. 3. Avoid including any explanations in the translation. 4.Please review the translations for verifying that they remain faithful to the original text and provide revised versions accordingly if necessary. If no revisions are needed, provide the translations as they are.

In our study, we conducted several experiments to evaluate the performance of GPT-3.5. The following were the approaches we employed:

- Sentence-to-sentence translation: We used the prompt "Translate this sentence from SRC to TGT, do not write any explanations" to evaluate the model's ability to translate individual sentences accurately.
- Multi-turn dialogue translation: We explored the impact of multi-turn conversations on the performance of GPT-3.5. Using the prompt P1.
- Multi-turn dialogue translation with detailed prompt P3. This experiment aims to test whether GPT-3.5 has the ability to get faithful translations while minimizing unaligned translations.
- Comparison between GPT-3.5-4k and GPT-3.5-16k: We performed separate experiments using both GPT-3.5-4k and GPT-3.5-16k models to observe any differences in translation abilities between the two.
- Adjusting temperature parameter: We varied the temperature parameter (0, 0.3, 0.7) to examine its impact on the translation quality. Changing the temperature can control the randomness of the generated translations.
- Incorporating fake CoT prompt P3. This experiment aims to test whether GPT-3.5 has the ability to automatically reflect and optimize its translations.

3 Result

We conduct experiments to quantify the impact of each component in our system. The evaluation conduct on test set on wmt22 using SacreBLEU (Post, 2018) and COMET (Stewart et al., 2020).

As shown in Table 1, here are the conclusions based on your experimental results:

- From the first and second experiment results, it can be concluded that the performance of GPT-3.5 in multi-turn dialogue is better than

language pair	Prompt	Multi-turn	T	Model	Bleu-A	Bleu-B	Chrf-A	Chrf-B	Comet-A	Comet-B
zh-en	P1	false	0	GPT-3.5-4k	26.6	20.0	57.4	52.5	52.7	43.5
zh-en	P1	true	0	GPT-3.5-4k	27.7	20.7	58.4	53.2	55.8	46.7
zh-en	P3	true	0	GPT-3.5-16k	23.4	18.0	54.4	50.2	54.9	46.1
en-zh	P1	true	0	GPT-3.5-4k	45.7	53.9	41.1	48.5	63.4	71.2
en-zh	P2	true	0	GPT-3.5-4k	44.2	51.4	39.9	46.0	62.1	70.6
en-zh	P2	true	0.7	GPT-3.5-4k	42.8	49.3	38.4	44.1	61.7	68.7
en-zh	P2	true	0.7	GPT-3.5-16k	42.7	49.3	38.3	44.8	63.2	71.1
en-zh	P2	true	0.3	GPT-3.5-16k	44.4	51.5	39.9	46.3	63.8	71.2

Table 1: Bleu/Chrf/Comet score on wmt22 test set. The COMET scores are calculated with the model wmt20-comet-da, the ChrF scores are calculated using all available references and SacreBLEU signature is the default settings. Scores are multiplied by 100. T represents Temperature

single-turn translation. This indicates that context can help improve the translation quality of GPT-3.5 by providing additional prompts.

- Comparing the results of the third experiment with the fourth experiment, it is concluded that the performance of P2 is worse. This suggests that GPT-3.5 does not fully understand the given prompt, which results in difficulty in generating accurate translations.

- Comparing the results of the fourth, fifth, and seventh experiments, it is concluded that lower temperature values yield better translation results. This indicates that reducing temperature parameter leads to more deterministic and high-quality translations.

- Comparing the results of the fifth and sixth experiments, it is concluded that GPT-3.5-16k performs better in translation than GPT-3.5-4k.

- Comparing the results of the seventh experiment with previous results, it is concluded that P3 performs the worst. Additionally, observing the actual revised results, it can be noted that GPT-3.5-16k rarely modifies its translations, indicating that without specific and clear instructions, it is unable to make effective modifications to its own translations.

Based on our previous results, we have chosen GPT-3.5-16k as the final model for our submission. For the WMT23 en-zh/zh-en track, we set the temperature to 0 and utilized P1 as the prompt. Adopting a multi-turn dialogue approach, we submitted our final results with the systerm name "LanBridgeMT". Figure 1 and Figure 2 show the results of our systerm. [1] Additionally, for other language pairs in the general WMT competition, we opted to submit the results generated by our LanMT (Han et al., 2022) engine. This decision was made to assess the engine's performance and determine its scoring capabilities directly in the online evaluation environment.

By taking these approaches, we aim to showcase the effectiveness of GPT-3.5 and demonstrate the performance of our LanMT engine in the respective WMT tracks. These submissions reflect our overarching goal of participating in and contributing to the advancement of machine translation research and development.

4 Conclusion

In this study, we evaluated the translation performance of GPT-3.5 using various experimental approaches. Our findings indicate that incorporating multi-turn dialogue prompts improves the translation quality of GPT-3.5, highlighting the importance of context in guiding the model's translations. Furthermore, we observed that GPT-3.5-16k, compared to GPT-3.5-4k, demonstrates superior translation capabilities in commit scores, indicating its enhanced ability to understand and fulfill user instructions. However, there are marginal differences in the other two metrics, BLEU and ChrF. Additionally, we found that lower temperature values result in improved translation quality, indicating the usefulness of controlling randomness in the generated translations. However, it is important to note that GPT-3.5 may struggle with understanding ambiguous prompts and lacks the ability to autonomously adjust and optimize its translations without explicit instructions. These findings contribute to our under-

[1] https://github.com/wmt-conference/wmt23-news-systems

System	COMET
HW-TSC	82.8
ONLINE-B	82.7
Yishu	82.7
GPT4-5shot	81.6
Lan-BridgeMT	81.2
ONLINE-G	80.9
ONLINE-Y	80.6
ONLINE-A	80.3
ZengHuiMT	79.6
ONLINE-W	79.3
IOL_Research	79.2
ONLINE-M	77.7
NLLB_MBR_BLEU	76.8
ANVITA	76.6
NLLB_Greedy	76.4

System	chrF
HW-TSC	57.5
ONLINE-B	57.5
Yishu	57.4
ZengHuiMT	54.6
ONLINE-G	53.9
ONLINE-A	53.4
GPT4-5shot	53.1
Lan-BridgeMT	53.1
ONLINE-W	52.5
IOL_Research	52.4
ONLINE-Y	52.3
ONLINE-M	49.7
ANVITA	47.1
NLLB_Greedy	46.1
NLLB_MBR_BLEU	45.8

System	BLEU
HW-TSC	33.6
ONLINE-B	33.5
Yishu	33.4
ONLINE-A	28.3
Lan-BridgeMT	27.3
IOL_Research	27.2
ZengHuiMT	27.0
GPT4-5shot	26.8
ONLINE-G	26.6
ONLINE-W	26.4
ONLINE-Y	25.0
ONLINE-M	23.5
ANVITA	21.8
NLLB_Greedy	20.5
NLLB_MBR_BLEU	19.8

Figure 1: Score for zh-en translation task

System	COMET
ONLINE-B	88.1
Yishu	88.1
HW-TSC	87.3
GPT4-5shot	87.1
ONLINE-W	86.8
Lan-BridgeMT	86.6
ONLINE-Y	86.5
ONLINE-A	86.2
IOL_Research	85.3
ZengHuiMT	84.3
ONLINE-M	84.2
ONLINE-G	83.8
NLLB_Greedy	75.7
ANVITA	75.6
NLLB_MBR_BLEU	71.5

System	chrF
HW-TSC	53.8
Yishu	53.0
ONLINE-B	52.9
ONLINE-A	52.8
IOL_Research	51.9
ONLINE-M	50.6
ONLINE-Y	49.8
ONLINE-G	49.4
ONLINE-W	47.3
ZengHuiMT	47.0
Lan-BridgeMT	46.8
GPT4-5shot	46.5
ANVITA	36.9
NLLB_Greedy	26.3
NLLB_MBR_BLEU	21.1

System	BLEU
HW-TSC	58.6
ONLINE-A	58.5
Yishu	57.6
ONLINE-B	57.5
IOL_Research	56.9
ONLINE-M	54.9
ONLINE-Y	54.2
ONLINE-G	54.1
ZengHuiMT	52.9
ONLINE-W	52.1
Lan-BridgeMT	50.2
GPT4-5shot	49.6
ANVITA	38.9
NLLB_Greedy	27.4
NLLB_MBR_BLEU	19.1

Figure 2: Score for en-zh translation task

standing of the strengths and limitations of GPT-3.5 in translation tasks, emphasizing the need for precise prompts to achieve optimal translation results.

References

Gpt-4 technical report.

Bing Han, Yangjian Wu, Gang Hu, and Qiulin Chen. 2022. Lan-bridge MT's participation in the WMT 2022 general translation shared task. In Proceedings of the Seventh Conference on Machine Translation (WMT), pages 268–274, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Craig Stewart, Ricardo Rei, Catarina Farinha, and Alon Lavie. 2020. COMET - deploying a new state-of-the-art MT evaluation metric in production. In Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track), pages 78–109, Virtual. Association for Machine Translation in the Americas.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems, 30.

Shuoheng Yang, Yuxin Wang, and Xiaowen Chu. 2020. A survey of deep learning techniques for neural machine translation. arXiv preprint arXiv:2002.07526.

Treating General MT Shared Task as a Multi-Domain Adaptation Problem: HW-TSC's Submission to the WMT23 General MT Shared Task

Zhanglin Wu, Daimeng Wei, Zongyao Li, Zhengzhe Yu, Shaojun Li, Xiaoyu Chen, Hengchao Shang, Jiaxin Guo, Yuhao Xie, Lizhi Lei, Hao Yang, Yanfei Jiang
Huawei Translation Service Center, Beijing, China
{wuzhanglin2,weidaimeng,lizongyao,yuzhengzhe,lishaojun18,chenxiaoyu35,
shanghengchao,guojiaxin1,xieyuhao2,leilizhi,yanghao30,jiangyanfei}@huawei.com

Abstract

This paper presents the submission of Huawei Translate Services Center (HW-TSC) to the WMT23 general machine translation (MT) shared task. We participate in Chinese↔English (zh↔en) language pair. We use deep Transformer architecture and obtain the best performance via a Transformer variant with a larger parameter size. We perform fine-grained pre-processing and filtering on the provided large-scale bilingual and monolingual datasets. The model enhancement strategies we used includes Regularized Dropout, Bidirectional Training, Data Diversification, Forward Translation, Back Translation, Alternated Training, Curriculum Learning and Transductive Ensemble Learning. Our submission obtain competitive results in the final evaluation.

1 Introduction

Machine translation (MT) (Brown et al., 1990) refers to the automatic translation of text from one language to another, while the WMT23 general MT shared task focuses on evaluation of general MT capabilities. Compared with the news shared task in previous years, the general MT shared task involves multiple domains. The testsets contain data in news, user generated (social), conversational, and ecommerce domains.

This paper presents the submission of HW-TSC to the WMT23 general MT shared task, in which we participate in zh↔en language pair. Our method is mainly based on previous works (Wei et al., 2022; Wu et al., 2022; Yang et al., 2021). We perform multi-step data cleansing on the provided dataset and only keep a high-quality subset for training. At the same time, several model enhancement strategies are tested in a pipeline, including Regularized Dropout (Wu et al., 2021), Bidirectional Training (Ding et al., 2021), Data Diversification (Nguyen et al., 2020), Forward Translation (Abdulmumin, 2021), Back Translation (Sennrich et al., 2016), Alternated Training (Jiao et al., 2021), Curriculum Learning (Zhang et al., 2019) and Transductive Ensemble Learning (Wang et al., 2020b).

Our system report includes four parts. Section 2 focuses on our data processing strategies while section 3 describes our training details. Section 4 explains our experiment settings and training processes and section 5 presents the results.

2 Data

2.1 Data Source

We obtain bilingual and monolingual data from ParaCrawl v9, News Commentary v18.1, Wiki Titles v3, UN Parallel Corpus V1.0, CCMT Corpus, WikiMatrix, News Crawl and Common Crawl data sources. The amount of data we used is shown in Table 1. It should be noted that in order to obtain better performance in the general domain, we mix the monolingual data from Common Crawl and News Crawl.

language pairs	bitext data	monolingual data
zh↔en	25M	en: 50M, zh: 50M

Table 1: Bilingual and monolingual used for training.

2.2 Data Pre-processing

Our data processing procedure is precisely the same as the previous year (Wei et al., 2021), including deduplication, XML content processing, langid (Lui and Baldwin, 2012) and fast-align (Dyer et al., 2013) filtering strategies. As we use the same data pre-processing strategy as the previous year, we will not go into details here.

2.3 Data Denoising

Since there may be some semantically dissimilar sentence pairs in bilingual data, we use LaBSE (Feng et al., 2022) to calculate the semantic similarity of each bilingual sentence pair, and exclude

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 170–174
December 6–7, 2023.

bilingual sentence pairs with a similarity score lower than 0.7 from our training corpus.

3 System Overview

3.1 Model

We continue using Transformer (Vaswani et al., 2017) as our neural machine translation (NMT) (Bahdanau et al., 2015) model architecture. As we did last year, we only use a 25-6 deep model architecture (Wang et al., 2019). The parameters of the model are the same as Transformer big. We just change the post-layer normalization to the pre-layer normalization, and set encoder layers to 25.

3.2 Regularized Dropout

Regularized Dropout (R-Drop)[1] (Wu et al., 2021) is a simple yet more effective alternative to regularize the training inconsistency induced by dropout (Srivastava et al., 2014). Concretely, in each mini-batch training, each data sample goes through the forward pass twice, and each pass is processed by a different sub model by randomly dropping out some hidden units. R-Drop forces the two distributions for the same data sample outputted by the two sub models to be consistent with each other, through minimizing the bidirectional Kullback-Leibler (KL) divergence (Van Erven and Harremos, 2014) between the two distributions. That is, R-Drop regularizes the outputs of two sub models randomly sampled from dropout for each data sample in training. In this way, the inconsistency between the training and inference stage can be alleviated.

3.3 Bidirectional Training

Many studies have shown that pre-training can transfer the knowledge and data distribution, hence improving the model generalization. Bidirectional training (BiT) (Ding et al., 2021) is a simple and effective pre-training method for NMT. Bidirectional training is divided into two stages: (1) bidirectionally updates model parameters, and (2) tune the model. To achieve bidirectional updating, we only need to reconstruct the training samples from "src→tgt" to "src→tgt & tgt→src" without any complicated model modifications. Notably, BiT does not require additional parameters or training steps and only uses parallel data.

[1] https://github.com/dropreg/R-Drop

3.4 Data Diversification

Data Diversification (DD) (Nguyen et al., 2020) is a data augmentation method to boost NMT performance. It diversifies the training data by using the predictions of multiple forward and backward models and then merging them with the original dataset which the final NMT model is trained on. DD is applicable to all NMT models. It does not require extra monolingual data, nor does it add more parameters. To conserve training resources, we only use one forward model and one backward model to diversify the training data.

3.5 Forward Translation

Forward translation (FT) (Abdulmumin, 2021), also known as self-training, is one of the most commonly used data augmentation methods. FT has proven effective for improving NMT performance by augmenting model training with synthetic parallel data. Generally, FT is performed in three steps: (1) randomly sample a subset from the large-scale source monolingual data; (2) use a "teacher" NMT model to translate the subset data into the target language to construct the synthetic parallel data; (3) combine the synthetic and authentic parallel data to train a "student" NMT model.

3.6 Back Translation

An effective method to improve NMT with target monolingual data is to augment the parallel training data with back translation (BT) (Sennrich et al., 2016; Wei et al., 2023). There are many works expand the understanding of BT and investigates a number of methods to generate synthetic source sentences. Edunov et al. (2018) find that back translations obtained via sampling or noised beam outputs are more effective than back translations generated by beam or greedy search in most scenarios. Caswell et al. (2019) show that the main role of such noised beam outputs is not to diversify the source side, but simply to tell the model that the given source is synthetic. Therefore, they propose a simpler alternative strategy: Tagged BT. This method uses an extra token to mark back translated source sentences, which generally outperforms noised BT (Edunov et al., 2018). For better joint use with FT, we use sampling back translation (ST) (Edunov et al., 2018).

3.7 Alternated Training

While synthetic bilingual data have demonstrated their effectiveness in NMT, adding more synthetic data often deteriorates translation performance since the synthetic data inevitably contains noise and erroneous translations. Alternated training (AT) (Jiao et al., 2021) introduce authentic data as guidance to prevent the training of NMT models from being disturbed by noisy synthetic data. AT describes the synthetic and authentic data as two types of different approximations for the distribution of infinite authentic data, and its basic idea is to alternate synthetic and authentic data iteratively during training until the model converges.

3.8 Curriculum Learning

A practical curriculum learning (CL) (Zhang et al., 2019) method should address two main questions: how to rank the training examples, and how to modify the sampling procedure based on this ranking. For ranking, we choose to estimate the difficulty of training samples according to their domain feature (Wang et al., 2020a). The calculation formula of domain feature is as follows, where θ_{in} represents an in-domain NMT model, and θ_{out} represents a out-of-domain NMT model. One thing to note is that we treat domains including news, user-generated (social), conversational, and e-commerce domains as in-domain, and others as out-of-domain. Specifically, we use the WMT22 test set to fine-tune a baseline model, and then use the baseline model and the fine-tuned model as the out-of-domain model and the in-domain model respectively.

$$q(x,y) = \frac{\log P(y|x;\theta_{in}) - \log P(y|x;\theta_{out})}{|y|} \quad (1)$$

For sampling, we adopt a probabilistic CL strategy that leverages the concept of CL in a nondeterministic fashion without discarding the original standard training practice, such as bucketing and mini-batching.

3.9 Transductive Ensemble Learning

Ensemble learning (Garmash and Monz, 2016), which aggregates multiple diverse models for inference, is a common practice to improve the performance of machine learning models. However, it has been observed that the conventional ensemble methods only bring marginal improvement for NMT when individual models are strong or there are a large number of individual models. Transductive Ensemble Learning (TEL) (Zhang et al., 2019) studies how to effectively aggregate multiple NMT models under the transductive setting where the source sentences of the test set are known. TEL uses all individual models to translate the source test set into the target language space and then fine-tune a strong model on the translated synthetic data, which significantly boosts strong individual models and benefits a lot from more individual models.

4 Experiment Settings

We use the open-source fairseq (Ott et al., 2019) for training, then we use SacreBLEU (Post, 2018)[2] and wmt20-comet-da model (Rei et al., 2020) to measure system performances. The main parameters are as follows: each model is trained using 8 A100 GPUs, batch size is 6144, parameter update frequency is 2, and learning rate is 5e-4. The number of warmup steps is 4000, and model is saved every 1000 steps. The architecture we used is described in section 3.1. We adopt dropout, and the rate varies across different training phases. R-Drop is used in model training, and we set λ to 5.

5 Results

Regarding zh↔en, we use Regularized Dropout, Bidirectional Training, Data Diversification, Forward Translation, Back Translation, Alternated Training, Curriculum Learning, and Transductive Ensemble Learning. The evaluation results of en→zh and zh→en NMT system on WMT22 general test sets are shown in Tables 2.

	en→zh		zh→en	
	BLEU	COMET	BLEU	COMET
BiT R-Drop baseline	45.55	50.24	22.30	22.28
+ DD, FT & ST	49.54	59.69	25.67	33.44
+ AT	54.11	63.99	28.58	37.15
+ CL	56.36	68.90	30.58	44.62
+ TEL	**56.80**	**69.06**	**31.35**	**45.56**

Table 2: BLEU and COMET scores of en→zh and zh→en NMT system on WMT22 general test set.

We observe that DD, FT & ST can stably bring 3-4 BLEU and 1-9 COMET improvement; AT can bring 3-5 BLEU and 4 COMET improvement; and CL can bring 2 BLEU and 5-7 COMET improvement. In addition, TEL can further slightly improve BLEU and COMET scores. Our final en→zh

[2] https://github.com/mjpost/sacrebleu

System	chrF	BLEU	COMET
HW-TSC	**57.5**	**33.6**	**82.8**
ONLINE-B	**57.5**	33.5	82.7
Yishu	57.4	33.4	82.7
GPT4-5shot	53.1	26.8	81.6
Lan-BridgeMT	53.1	27.3	81.2
ONLINE-G	53.9	26.6	80.9
ONLINE-Y	52.3	25.0	80.6
ONLINE-A	53.4	28.3	80.3
ZengHuiMT	54.6	27.0	79.6
ONLINE-W	52.5	26.4	79.3
IOL_Research	52.4	27.2	79.2
ONLINE-M	49.7	23.5	77.7
NLLB_MBR_BLEU	45.8	19.8	76.8
ANVITA	47.1	21.8	76.6
NLLB_Greedy	46.1	20.5	76.4

Table 3: Scores for the WMT23 zh→en translation task: chrF, BLEU and COMET (Unbabel/wmt22-comet-da).

System	chrF	BLEU	COMET
ONLINE-B	52.9	57.5	**88.1**
Yishu	53.0	57.6	**88.1**
HW-TSC	**53.8**	**58.6**	87.3
GPT4-5shot	46.5	49.6	87.1
ONLINE-W	47.3	52.1	86.8
Lan-BridgeMT	46.8	50.2	86.6
ONLINE-Y	49.8	54.2	86.5
ONLINE-A	52.8	58.5	86.2
IOL_Research	51.9	56.9	85.3
ZengHuiMT	47.0	52.9	84.3
ONLINE-M	50.6	54.9	84.2
ONLINE-G	49.4	54.1	83.8
NLLB_Greedy	26.3	27.4	75.7
ANVITA	36.9	38.9	75.6
NLLB_MBR_BLEU	21.1	19.1	71.5

Table 4: Scores for the WMT23 en→zh translation task: chrF, BLEU, COMET (Unbabel/wmt22-comet-da).

and zh→en submissions achieve 56.80 and 31.35 BLEU, 69.06 and 45.56 COMET respectively.

6 Official Automatic Evaluation Results

In our final submission, we add post-processing for punctuation correction and entity preservation. WMT (Kocmi et al., 2023) present an automatic evaluation of the systems submitted to the general machine translation task, including the following three different automatic metrics: chrF, BLEU and COMET. We rank the systems according to COMET scores, and unconstrained systems are in a grey background in the tables.

7 Conclusion

This paper presents the submission of HW-TSC to the WMT23 general MT Task. We participate in zh↔en language pair and perform experiments with a series of pre-processing and training strategies. The effectiveness of each strategy is demonstrated. Our experiments show that our model training strategies are effective. Our submission finally achieve competitive results in the evaluation.

References

Idris Abdulmumin. 2021. Enhanced back-translation for low resource neural machine translation using self-training. In *Information and Communication Technology and Applications: Third International Conference, ICTA 2020, Minna, Nigeria, November 24–27, 2020, Revised Selected Papers*, volume 1350, page 355. Springer Nature.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, John Lafferty, Robert L Mercer, and Paul S Roossin. 1990. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63.

Liang Ding, Di Wu, and Dacheng Tao. 2021. Improving neural machine translation by bidirectional training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3278–3284.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, page 489. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.

Ekaterina Garmash and Christof Monz. 2016. Ensemble learning for multi-source neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1409–1418.

Rui Jiao, Zonghan Yang, Maosong Sun, and Yang Liu. 2021. Alternated training with synthetic and authentic data for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1828–1834.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and Maja Popović. 2023. Findings of the 2023 conference on machine translation (WMT23). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30.

Xuan-Phi Nguyen, Shafiq Joty, Wu Kui, and Ai Ti Aw. 2020. Data diversification: a simple strategy for neural machine translation. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 10018–10029.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96. Association for Computational Linguistics (ACL).

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Tim Van Erven and Peter Harremos. 2014. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. 2019. Learning deep transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Wei Wang, Ye Tian, Jiquan Ngiam, Yinfei Yang, Isaac Caswell, and Zarana Parekh. 2020a. Learning a multi-domain curriculum for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7711–7723.

Yiren Wang, Lijun Wu, Yingce Xia, Tao Qin, ChengXiang Zhai, and Tie-Yan Liu. 2020b. Transductive ensemble learning for neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6291–6298.

Daimeng Wei, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Xiaoyu Chen, Hengchao Shang, Jiaxin Guo, Minghan Wang, Lizhi Lei, Min Zhang, et al. 2021. Hw-tsc's participation in the wmt 2021 news translation shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 225–231.

Daimeng Wei, Zhiqiang Rao, Zhanglin Wu, Shaojun Li, Yuanchang Luo, Yuhao Xie, Xiaoyu Chen, Hengchao Shang, Zongyao Li, Zhengzhe Yu, et al. 2022. Hw-tsc's submissions to the wmt 2022 general machine translation shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 403–410.

Daimeng Wei, Zhanglin Wu, Hengchao Shang, Zongyao Li, Minghan Wang, Jiaxin Guo, Xiaoyu Chen, Zhengzhe Yu, and Hao Yang. 2023. Text style transfer back-translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 7944–7959, Toronto, Canada.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34:10890–10905.

Zhanglin Wu, Jinlong Yang, Zhiqiang Rao, Zhengzhe Yu, Daimeng Wei, Xiaoyu Chen, Zongyao Li, Hengchao Shang, Shaojun Li, Ming Zhu, et al. 2022. Hw-tsc translation systems for the wmt22 biomedical translation task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 936–942.

Hao Yang, Zhanglin Wu, Zhengzhe Yu, Xiaoyu Chen, Daimeng Wei, Zongyao Li, Hengchao Shang, Minghan Wang, Jiaxin Guo, Lizhi Lei, et al. 2021. Hw-tsc's submissions to the wmt21 biomedical translation task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 879–884.

Xuan Zhang, Pamela Shapiro, Gaurav Kumar, Paul McNamee, Marine Carpuat, and Kevin Duh. 2019. Curriculum learning for domain adaptation in neural machine translation. In *Proceedings of NAACL-HLT*, pages 1903–1915.

UvA-MT's Participation in the WMT23 General Translation Shared Task

Di Wu* **Shaomu Tan*** **David Stap** **Ali Araabi** **Christof Monz**
Language Technology Lab
University of Amsterdam
{d.wu, s.tan, d.stap, a.araabi, c.monz}@uva.nl

Abstract

This paper describes the UvA-MT's submission to the WMT 2023 shared task on general machine translation. We participate in the constrained track in two directions: English ↔ Hebrew. In this competition, we show that by using one model to handle bidirectional tasks, as a minimal setting of Multilingual Machine Translation (MMT), it is possible to achieve comparable results with that of traditional bilingual translation for both directions. By including effective strategies, like back-translation, re-parameterized embedding table, and task-oriented fine-tuning, we obtained competitive final results in the automatic evaluation for both English → Hebrew and Hebrew → English directions.

1 Introduction

Multilingual Machine Translation (MMT) (Johnson et al., 2017) has attracted a lot of attention in recent years because of 1) its high-level efficiency (multiple translation directions within a single model) and 2) the potential for knowledge transfer, especially for low-resource or even unseen directions. In MMT systems, only one additional tag is introduced to indicate the translation direction, compared to the conventional encoder-decoder architecture. In this competition, we explore MMT with a minimal setting, i.e., using one model for bidirectional translations simultaneously.

We leverage all the official parallel data and a substantial portion of the monolingual data generously provided by the WMT23 organizer, as elaborated in Section 2. To elevate the quality of our parallel data, we implemented a comprehensive three-step cleaning procedure. Additionally, for monolingual data, we further trained an n-gram language model to filter out low-quality sentences, with the goal of generating synthetic data as elaborated in Section 4.1.

The backbone of our system is based on a standard transformer (Vaswani et al., 2017). Additionally, we build a re-parameterized embedding table (Wu and Monz, 2023) (see Section 3.2) to enhance the representational word similarity between English and Hebrew, targeting better knowledge transfer for multilingual translation.

The final system involves three stages of training: 1) **Pretraining with synthetic data** (see Section 4.1), where we leverage back-translation (Sennrich et al., 2016) to produce synthetic data, and finally add them as additional data within new translation directions in our MMT system to conduct pretraining. 2) **Training without synthetic data** (see Section 4.2), where we discard the additional synthetic data and further train our system using real bitext only. 3) **Fine-tuning with task-related data** (see Section 4.3), where we copy and fine-tune our system using English → Hebrew and Hebrew → English data for each track respectively. We observe evident improvements for stage 1 and stage 2, while surprisingly performance drops for stage 3.

We report our results, including the offline evaluation and the final online evaluation, in Section 5. Our *constrained* system showed comparable performance to *unconstrained* systems, and outperformed the second-place constrained submission with +10 BLEU.

2 Data

In this section, we provide an overview of our data sources and the data cleaning procedures applied to our English-Hebrew translation task. We utilize both parallel and monolingual data sets provided by the organizers for training our translation systems.

2.1 Parallel Data

We make use of all the available data from the constrained track of the shared task for English-

*Equal contribution.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 175–180
December 6–7, 2023.

Hebrew translation. To enhance the quality of our parallel data, we undergo a thorough preprocessing phase involving three key steps, as outlined below. All steps in the cleaning step 1 are executed using the Moses toolkit[*] (Koehn et al., 2007). Consequently, we reduced the size of the raw bitext data from 70 million to 34 million sentences after completing the three steps of the cleaning process:

- Cleaning Step 1
  - Deescaping special characters in XML.
  - Removing non-printable characters.
  - Normalizing punctuation and tokenizing sentences using Moses.
- Cleaning Step 2
  - Filtering out sentences longer than 256 tokens.
  - Eliminating sentences where over 75% of the words on both the source and target sides are identical.
  - Removing sentences with a source-to-target token ratio exceeding 1.5.
  - Eliminating duplicate sentences.
- Cleaning Step 3
  - Removing off-target sentences using the FastText Language identification tool (Joulin et al., 2016).
  - Excluding sentences exhibiting one-to-many or many-to-one mappings, for example, a single source sentence having multiple different target sentences.

Furthermore, we sampled 10 million parallel sentences to learn a 32k joint unigram (Kudo, 2018) model-based subword vocabulary using SentencePiece (Kudo and Richardson, 2018), which we then utilized across all our models, including the n-gram KenLM model discussed in the next section. However, we encountered a situation where certain emoji tokens were not included in our vocabulary. As a result, we integrated an additional post-processing step in Section 4.4 to address this issue.

2.2 Monolingual Data

To enhance our translation systems further, we incorporate monolingual data to produce synthetic data through back-translation. For our monolingual data, we primarily rely on the official English data provided by the organizers. Note that we did not use any Hebrew monolingual data since it is limited (only 1 million sentences). We combine three official English monolingual datasets: News Discussions 2019, Leipzig News Corpora 2020, and News Crawl (2007-2022) to construct our raw monolingual dataset. Following this, we apply the same Cleaning Step 1 procedure as detailed in the Parallel Data section to preprocess the monolingual data, and this results in 373 million sentences.

Considering the low quality of monolingual data, we additionally filter them by training an n-gram language model, i.e., KenLM (Heafield, 2011), and eliminate the sentences below an LM score threshold. The training data of KenLM is all of the test data in English, including our offline test data Flores, and the official test dataset. We train KenLM at the subword level, where we use the same unigram model (trained upon original bilingual data) to split the training data of KenLM. Then, we use it to score all of the monolingual data. To establish a filtering threshold, we randomly selected 1,000 sentences and labeled them as positive or negative based on criteria such as fluency, naturalness (e.g., avoiding strings of numbers), and relevance to the domain mentioned for WMT23 test datasets. Finally, we chose a threshold that could filter 70% bad cases within the 1,000 sentences, with the cost of monolingual 30% data, resulting in around 250M total sentences.

Lastly, considering the limitation of the computational resource, we sample 32M monolingual sentences (at the same level as the bilingual dataset) from the filtered dataset.

3 Systems

3.1 Backbone and Baseline

In this section, we outline the foundational architecture and adjustments made to our baseline systems. Our baseline model leverages English $\leftrightarrow$ Hebrew translation directions by incorporating the target language token at the beginning of the encoder, denoted as "2he" and "2en". Our implementations are grounded in the Transformer architecture (Vaswani et al., 2017), leveraging the Fairseq toolkit (Ott et al., 2019).

For our baseline model, we utilize a 12-layer Transformer architecture (mT-large) with specific modifications, including pre-norm for both the en-

[*] https://github.com/moses-smt/mosesdecoder/

coder and decoder, and layer-norm for embedding. To enhance stability and performance, we tie the parameters of encoder embedding, decoder embedding, and decoder output. We also introduce dropout and attention dropout with a probability of 0.1, along with label smoothing at a rate of 0.1.

Similar to the approach described by Vaswani et al. (2017), we employ the Adam optimizer with a learning rate of 5e-4, implementing an inverse square root learning rate schedule with 4,000 warmup steps. We set the maximum number of tokens to 10,240, with gradient accumulation every 21 steps to facilitate large-batch training in Tang et al. (2021). We train all of our systems with 4 NVIDIA A6000 Gpus, and to expedite the training process, we conducted all experiments using half-precision training (FP16). Additionally, we save checkpoints every 2000 steps and implement early stopping based on perplexity, with a patience of 5 epochs.

3.2 Re-parameterized Embedding Table

Using a vocabulary that is shared across languages is common practice in MMT. In addition to its simple design, shared tokens play an important role in positive knowledge transfer, assuming that shared tokens refer to similar meanings across languages. This point has been demonstrated by previous works (Pires et al., 2019; Sun et al., 2022; Stap et al., 2023; Wu and Monz, 2023). To enhance word-level knowledge transfer, we follow (Wu and Monz, 2023) to implement a re-parameterized shared embedding table and equipped it with our backbone.

We leverage eflomal (Östling and Tiedemann, 2016) to train and extract subword-level alignments based on all of the bilingual data we used. Then, we build the priors of word equivalence (word alignments) into a graph and leverage GNN (Welling and Kipf, 2016) to re-parameterize the embedding table.

More specifically, For two words v_i and v_j in V, we define an alignment probability from v_j to v_i in corpus D as corresponding transfer ratios $g_{i,j}$ as follows:

$$g_{i,j} = \frac{c_{i,j}}{\sum_{k=1}^{|V|} c_{i,k}}, \quad (1)$$

where $c_{i,j}$ is the number of times both words are aligned with each other across D. The corresponding bilingual equivalence graph G can be induced by filling an adjacency matrix using $g_{i,j}$, G is applied within graph networks to re-parameterize the original embedding table as follows:

$$E' = \rho(EW_1 + GEW_2 + B). \quad (2)$$

To allow the message to pass over multiple hops, we stack multiple graph networks and calculate representations recursively as follows:

$$E^{h+1} = \rho(E^h W_1^h + GE^h W_2^h + B^h), \quad (3)$$

where h is the layer index, i.e., *hop*, and E^0 is equal to the original embedding table E. The last layer representation E^H is the final re-parameterized embedding table, for the maximum number of hops H, which is then used by the system just like any vanilla embedding table.

4 Experiments

We describe the training process of our system in three stages.

4.1 Pretraining with Synthetic Data

Back-translation plays an important role in leveraging monolingual data in machine translation. In this competition, we also apply it to produce synthetic data and include it in our first-stage training.

Specifically, we first train a base MMT model (backbone with re-parameterized embedding tables) using bilingual data. Then, we feed our monolingual English data to produce EN-HE synthetic bitext. Finally, we merge the original bilingual data with the synthetic data together to pre-train our MMT system. We follow Fan et al. (2021) and add an additional language tag "2syn" to differentiate between synthetic and original Hebrew data. Note that, although normally original data (here, it is EN) is used as target side data after back translation, we use synthetic data for both directions.

4.2 Training without Synthetic Data

Considering that the synthetic data may differ from the original bilingual data in terms of data quality, domain difference, and diversity, in the second-stage training, we encourage our system to skew towards the original bilingual distribution. We achieve this by discarding the synthetic data directly and continuing training upon the first-stage system as a kind of full parameter finetuning.

Strategy	Sampled Data		Full Data	
	EN→HE	HE→EN	EN→HE	HE→EN
Bilingual Baseline	24.6	31.1	34.1	46.0
MMT Baseline	24.7	31.6	34.1	45.8
MMT + GM 1-hop	**26.2**	32.3	**34.3**	**46.2**
MMT + GM 2-hop	25.5	**32.7**	-	-

Table 1: Offline evaluation results on sampled and full training data. For sampled data (2M), the backbone is Transformer Base, while for full data (34M) the backbone is Transformer Large as we describe in Section 3.1. MMT + GM means that we equip graph-based re-parameterized embedding tables for our MMT baseline, and hop means how many graph network layers are involved. The best BLEU scores in each column are written in bold.

Strategy	Offline		Online	
	EN→HE	HE→EN	EN→HE	HE→EN
MMT Baseline	34.1	45.8	33.3	50.3
MMT + GM 1-hop	34.3	46.2	33.6	50.7
MMT + GM 1-hop + Stage-1	**35.4**	46.8	**35.0**	50.1
MMT + GM 1-hop + Stage-1,2	34.1	**47.4**	**35.0**	**51.0**
MMT + GM 1-hop + Stage-1,2,3	33.3	44.3	33.3	48.0

Table 2: Final results of three stages training. The best BLEU scores in each column are written in bold.

4.3 Finetuning on Task-Specific Data

Lastly, to encourage the system to focus on one certain language direction, we further fine-tune direction-specific data on the second-stage system. Note, the direction-specific data here, i.e., EN → HE and HE → EN are both from the original bilingual data. The effectiveness is also demonstrated by Ding et al. (2021); Zan et al. (2022) for bilingual translation.

In short, in this three-stage training process, we gradually narrow down the data distribution to focus on task-specific real data.

4.4 Post-Processing

We noticed that some emoji tokens in the official test set were not included in our vocabulary. Thus, we integrated an additional post-processing step to process them. Specifically, we escaped the emoji tokens to their Unicode string* before tokenizing and feeding them to our system to conduct inference, and then convert the Unicode string back for generated predictions.

4.5 Offline Evaluation

We used Flores-200 (Costa-jussà et al., 2022) to evaluate our strategies offline before submissions and Ntrex-128 (Federmann et al., 2022) as the validation set. We show the results in Table 1. Due to resource limitations, we sample 2M of bilingual data to verify whether there is a big performance gap between MMT and bilingual baseline. Meanwhile, we also chose the best hyperparameter for our re-parameterized embedding table, i.e., the hop number, based on the sampled dataset.

As shown in Table 1, on both sampled and full data, the MMT baseline achieves comparable results with bilingual counterparts. Especially for sampled data, it even outperforms 0.5 BLEU for into-English translation.

The model-equipped 1-hop re-parameterized embedding table demonstrates a notable improvement, yielding a 1.5 BLEU gain for the out-of-English direction and a 0.7 BLEU gain for the out-of- and into-English directions, on 2M datasets. It shows that the embedding re-parameterized method (Wu and Monz, 2023) also works for bilingual settings, which is not explored in the original paper. We did not observe evident gains for 2-hop compared with 1-hop on sampled data, hence, we only apply the 1-hop graph networks for the full data training. As shown in the table, the results are consistent with that of small data, where MMT with 1-hop graph networks achieve better performance than MMT baseline.

As above, we chose MMT with the 1-hop setting as our architecture and conducted our three stages of training as described in Section 4.

*For example, the emoji of "Grinning Face with Open Eyes" will convert to a string "U+1F600".

5 Results

Table-2 shows our offline and online evaluation results according to each training stage described in Section 4. We still use Flores-200 to conduct offline evaluations. The online results are reported by WMT23 background BLEU evaluations. Stage-1, -2, and -3 refer to "Pretraining with Synthetic Data", "Training without Synthetic Data", and "Finetuning on Task-Specific Data" respectively.

The results of online and offline evaluations are quite consistent. Both of them achieve best results when training with stage 1 and stage 2. It shows that by step-by-step narrowing training data from mixing with synthetic data to real data distribution, we can further boost our MMT system's performance. However, when we further conduct fine-tuning on direction-specific data, i.e., applying stage 3, there is an evident performance drop. It seems that tuning in a specific direction upon MMT may not be a good practice, at least when the training data are a subset of that for MMT. We leave this point for future exploration.

Our final system achieves 35.0 and 51.0 in EN→HE and HE→EN direction respectively, which are both in the first place for constrained tracks.

6 Conclusion

In this competition, we show that: 1) It is possible to achieve comparable results with conventional bilingual translation by using MMT training fashion to handle two dual translation directions. 2) Previous embedding re-parameterized method (Wu and Monz, 2023) also works for bilingual translation, which is not verified in the original paper. However, when training data scales up to 30+M level, the improvements become marginal. 3) By step-by-step narrowing training data (especially for stage-1 and stage-2) from mixing with synthetic data to real data distribution, we successfully boost the final performance, even in a quite high-resource scenario (30+M).

Acknowledgments

This research was funded in part by the Netherlands Organization for Scientific Research (NWO) under project number VI.C.192.080.

References

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Liang Ding, Di Wu, and Dacheng Tao. 2021. Improving neural machine translation by bidirectional training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3278–3284, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1):4839–4886.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. Ntrex-128–news test references for mt evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with markov chain monte carlo. *The Prague Bulletin of Mathematical Linguistics*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

David Stap, Vlad Niculae, and Christof Monz. 2023. Viewing knowledge transfer in multilingual machine translation through a representational lens. *arXiv preprint arXiv:2305.11550*.

Simeng Sun, Angela Fan, James Cross, Vishrav Chaudhary, Chau Tran, Philipp Koehn, and Francisco Guzmán. 2022. Alternative input signals ease transfer in multilingual machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5291–5305, Dublin, Ireland. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Max Welling and Thomas N Kipf. 2016. Semi-supervised classification with graph convolutional networks. In *J. International Conference on Learning Representations (ICLR 2017)*.

Di Wu and Christof Monz. 2023. Beyond shared vocabulary: Increasing representational word similarities across languages for multilingual machine translation. *arXiv preprint arXiv:2305.14189*.

Changtong Zan, Keqin Peng, Liang Ding, Baopu Qiu, Boan Liu, Shwai He, Qingyu Lu, Zheng Zhang, Chuang Liu, Weifeng Liu, Yibing Zhan, and Dacheng Tao. 2022. Vega-MT: The JD explore academy machine translation system for WMT22. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 411–422, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Achieving State-of-the-Art Multilingual Translation Model with Minimal Data and Parameters

Hui Zeng
LanguageX AI Lab
felix_zeng_ai@aliyun.com

Abstract

This is LanguageX (ZengHuiMT)'s submission to WMT 2023 General Machine Translation task for 13 language directions. We initially employ an encoder-decoder model to train on all 13 competition translation directions as our baseline system. Subsequently, we adopt a decoder-only architecture and fine-tune a multilingual language model by partially sampling data from diverse multilingual datasets such as CC100 and WuDaoCorpora. This is further refined using carefully curated high-quality parallel corpora across multiple translation directions to enable the model to perform translation tasks. As per automated evaluation metrics, our model ranks first in the translation directions from English to Russian, English to German, and English to Ukrainian. It secures the second position in the directions from English to Czech, English to Hebrew, Hebrew to English, and Ukrainian to English, and ranks third in German to English, Japanese to English, and Russian to English among all participating teams. Our best-performing model, covering 13 translation directions, stands on par with GPT-4. Among all 13 translation directions, our multilingual model surpasses GPT-4 in bleu scores for 7 translation directions.

1 Introduction

Since 2023, large language models like ChatGPT (Brockman et al., 2023) have had a profound impact on the field of machine translation, characterized by an ever-increasing scale in terms of parameters and data requirements. Many research institutions and language service providers struggle to keep pace with this computational arms race. For smaller teams, the only viable strategy is to maximize model performance under constrained hardware resources. We participated in the WMT 2023 General Machine Translation task, covering 13 translation directions. Given our limited computational power and time constraints, it was infeasible to craft dedicated models for each translation direction, making a large-scale multilingual translation model our optimal choice. We utilized Fairseq (Ott et al., 2019) to train our baseline multilingual translation model and further employed the Hugging Face Transformers Toolkit (Wolf et al., 2020) to train a multilingual language model. Subsequent fine-tuning with task-specific instructions enabled it to perform multilingual translation tasks effectively.

2 Data Filtering and Selection

We participated in the WMT 2023 General MT task, competing in 13 language pairs including Chinese to/from English, German to/from English (at the document level), Hebrew to/from English (in a low-resource setting), Japanese to/from English, Russian to/from English, Ukrainian to/from English, and English to Czech.

Given the challenge of concurrently training for 13 translation directions, it was imperative for us to judiciously regulate the size of parallel corpora specific to each direction, as well as the parameter count of the multilingual translation model. This was crucial to ensure training completion within a constrained timeline. For the Chinese-English bi-directional translation, our primary sources for parallel corpora were the CCMT Corpus (which can be found at:

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 181–186

http://mteval.cipsc.org.cn:81/agreement/description), genuine internal translation project data, in addition to content extracted and curated from websites and e-books. This rigorous process resulted in a refined collection of approximately 5 million parallel sentence pairs. For other translation orientations, we used the English segments from the derived Chinese-English parallel corpus as foundational data. This seed data enabled the retrieval of analogous language pairs from our comprehensive in-house multilingual parallel corpus, with each translation direction maintaining a parallel sentence count in the ballpark of 5 million.

Given the need to train a decoder-only multilingual model, we primarily utilized public datasets such as Book Corpus (Zhu et al., 2015), CC100 (Conneau et al., 2020), and WuDaoCorpora (Yuan et al., 2021). It was also imperative for us to regulate the data volume for each language and the parameter count of the multilingual model. Table 1 delineates the sources and the respective quantities for monolingual data across different languages. Owing to the extended length of text segments in the WuDaoCorpora (Yuan et al., 2021), the number of extracted text passages is fewer compared to other languages. However, the character count remains substantial.

Language	Data Source	Size in GB	Paragraph Count
Chinese	WuDaoCorpora	6.4	3,980,000
Czech	CC100	4.5	29,630,985
German	CC100	5.1	24,958,540
English	BookCorpus	4.3	20,000,000
English	CC100	3.3	20,000,000
Hebrew	CC100	5.4	30,877,445
Japanese	CC100	5.8	30,985,700
Russian	CC100	5.7	13,928,244
Ukrainian	CC100	4.6	16,818,862

Table 1: Sources and Quantities of Monolingual Data for Each Language.

2.1 Monolingual Data Filtering

The following rules are used to filter parallel corpus.

• Remove duplicated sentence pairs.
• Remove the sentence pairs containing special characters.
• Remove the sentence pairs containing html addresses or tags.

2.2 Parallel Data Filtering Using Rules

The following rules are used to filter parallel corpus.

a. Remove duplicated sentence pairs.
b. Remove the lines having identical source and target sentences.
c. Remove the sentence pairs containing special characters.
d. Remove the sentence pairs containing html addresses or tags.
e. Remove the sentence pairs with empty source or target side.

2.3 Parallel Data Filtering Using Multilingual Language Model

We used a multilingual model - sentence-transformers/paraphrase-multilingual-mpnet-base-v2 (Reimers et al., 2019) that generates embeddings for sentences or paragraphs in various languages. Using these embeddings, we calculated semantic similarity scores for parallel sentence pairs. Based on these scores, we filtered out low-quality parallel sentence pairs.

3 System Description

This section illustrates how the model is trained step by step.

3.1 Data pre-processing

Data pre-processing of multilingual translation model. We utilized the NLLB (NLLB Team, 2022) tokenizer from Hugging Face as the foundation and incorporated additional Chinese tokens to create an enhanced tokenizer specifically for Chinese language processing. This resulted in a final vocabulary size of 266,786 tokens.

To ensure synchronized training across all translation directions and to prevent the model from mastering one translation direction at the expense of another, we evenly blended the multilingual parallel corpora. This involved sequentially placing a fixed number of parallel sentence pairs from different translation directions into the training set, typically set to 100 pairs per direction.

To facilitate the simultaneous training of multiple translation directions within a single large model, we shared the embeddings and vocabulary for both source and target languages. Furthermore, we prefixed the source part of the parallel sentence pairs with specific prompt tokens.

The structure of the parallel sentence pairs is as follows: {engine name} engine. Translation from {source language} to {target language}: {source line} 🐯 {target line} <eos>. 🐯 is the delimiter used for parallel corpora.

To better accommodate the German to/from English (at the document level) translation task, we combined conventional sentence-level German to/from English parallel corpora into paragraph-level corpora based on a specified number of sentences. We then mixed this with the regular sentence-level parallel corpora, ensuring the resultant model is trained to handle a broader range of sequence lengths.

Data pre-processing of multilingual language model. We employed the same tokenizer as used in the multilingual translation model.

Due to the vast size of the CC100 dataset (Conneau et al., 2020), we performed sampling on the data for all languages, with 1,000 lines as the sampling unit. Multiple units were extracted from various parts of the entire dataset to cover it as comprehensively as possible, while keeping the individual language data size at around 5GB.

To mitigate catastrophic forgetting, we uniformly mixed the monolingual data of each language. This ensured that the training process included synchronized training on data from all languages, rather than training on one language first and then training on another.

The structure of the supervised finetuning prompt for translation task is as follows: {engine name} engine. Text in {source language}: {source line} Translation of the previous text to {target language}: {target line} 🥰.

To prevent endless generation and excessive translation, the 🥰 emoji is placed at the end of the translation to signify its completion, signaling the model to cease generation.

3.2 Baseline Translation Model Training

The parallel data prepared in step 3.1 is used to train a multilingual translation model using transformer (Vaswani et al., 2017) architecture as the baseline. Training was conducted using Fairseq (Ott et al., 2019) over the entire dataset for four epochs. The crucial training parameters are as follows:

```
--encoder-layers 12 \
--encoder-attention-heads 16 \
--encoder-embed-dim 1024 \
--encoder-ffn-embed-dim 4096 \
--decoder-layers 6 \
--decoder-attention-heads 16 \
--decoder-embed-dim 1024 \
--decoder-ffn-embed-dim 4096 \
--share-decoder-input-output-embed \
--share-all-embeddings \
--max-source-positions 1024 \
--max-target-positions 1024 \
--lr 5e-4 \
--lr-scheduler inverse_sqrt \
--warmup-updates 4000 \
```

Parameter	Value
Trainable parameters	1,091,315,712
Vocabulary size	266,786
Max length	1024
Embedding Dimension	1536
Decoder layers	24
Attention heads	16
Learning rate	5e-5
Lr scheduler type	linear
Warmup steps	4,000

Table 1: Parameters for Training Multilingual Language Model.

Translation Direction	Baseline Translation Model	Multilingual Language Model
en-cs	41.20	**43.67**
en-de	40.20	**41.00**
en-he	35.00	**36.52**
en-ru	31.20	**32.07**
en-uk	26.60	**28.29**
en-zh	47.30	**53.01**
en-ja	17.00	**17.60**
de-en	26.70	**42.08**
he-en	56.00	**57.51**
ja-en	21.20	**23.54**
ru-en	30.90	**32.15**
uk-en	42.50	**44.28**
zh-en	25.2	**28.27**

Table 3: BLEU scores on Newstest 2023 for all directions and different training methodologies.

Translation Direction	GPT 4-5shot	Multilingual Language Model
en-cs	38.26	**43.67**
en-de	**44.08**	41.00
en-he	27.08	**36.52**
en-ru	31.09	**32.07**
en-uk	25.78	**28.29**
en-zh	49.65	**53.01**
en-ja	**20.55**	17.60
de-en	**49.54**	42.08
he-en	52.04	**57.51**
ja-en	**25.27**	23.54
ru-en	**35.31**	32.15
uk-en	**44.84**	44.28
zh-en	27.87	**28.27**

Table 4: BLEU score comparison between the multilingual model and GPT-4 across all language directions on Newstest 2023.

3.3 Multilingual Language Model Training

We utilized DeepSpeed (Rasley et al., 2020) and Hugging Face transformers (Vaswani et al., 2017) as our training tools and trained the models on the uniformly mixed monolingual data and SFT data prepared in step 3.1 after applying bf16 precision. The specific training parameters are presented in Table 2. The entire training process was completed using four RTX A6000 GPUs. After completing one full pass of the entire dataset, we

terminated the training of the multilingual language model.

3.4 Results

The BLEU (Papineni et al., 2002) scores on Newstest 2023 for all translation directions and different training methodologies are presented in Table 3.

Based on the automated assessment metrics, our system takes the lead in translation directions from English to Russian, English to German, and English to Ukrainian. It claims the runner-up spot for English to Czech, English to Hebrew, Hebrew to English, and Ukrainian to English directions, and occupies the third place for the German to English, Japanese to English, and Russian to English directions among the contenders.

4 Conclusion

This paper describes LanguageX (ZengHuiMT)'s translation system for the WMT2023 General MT task. Initially, we utilize a comprehensive encoder-decoder structure to establish our baseline system by training across all 13 contest translation directions. In the subsequent stages, we embrace a solely decoder-focused design and harness a multilingual language model, drawing samples from multilingual datasets like CC100 (Conneau et al., 2020) and WuDaoCorpora (Yuan et al., 2021). This model is then meticulously fine-tuned using select high-grade parallel corpora from various translation domains, empowering it to execute translation task.

Our best-performing model, covering 13 translation directions, boasts around 1 billion parameters. This is less than one percent of the parameter count of mammoth models like GPT-4 (OpenAI, 2023), which possess hundreds of billions of parameters. In translation evaluations across all languages, our system stands on par with GPT-4 (OpenAI, 2023). Among all 13 translation directions, our multilingual model surpasses GPT-4 (OpenAI, 2023) in bleu (Papineni et al., 2002) scores for 7 translation directions.

Acknowledgments

Thanks to my wife who spends most of her time to take care of our two kids, so that I am able to participate in the contest and complete this paper.

References

Greg Brockman, Atty Eleti, Elie Georges, Joanne Jang, Logan Kilpatrick, Rachel Lim, Luke Miller, and Michelle Pokrass. 2023. Introducing ChatGPT and Whisper APIs. *https://openai.com/blog/introducing-chatgpt-and-whisper-apis*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *In Proceedings of the 31st International Conference on Neural Information Processing Systems,* pages 6000–6010.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wen-zek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov: Unsupervised Cross-lingual Representation Learning at Scale. *In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451. (2020)

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Wolf and Lysandre Debut and Victor Sanh and Julien Chaumond and Clement Delangue and Anthony Moi and Pierric Cistac and Tim Rault and Rémi Louf and Morgan Funtowicz and Joe Davison and Sam Shleifer and Patrick von Platen and Clara Ma and Ya-cine Jernite and Julien Plu and Canwen Xu and Teven Le Scao and Sylvain Gugger and Ma-riama Drame and Quentin Lhoest and Alexander M. Rush: Transformers: State-of-the-Art Natural Language Processing. *In: Proceedings of the 2020 Conference on Empirical Meth-ods in Natural Language Processing: System Demonstrations,* pp. 38–45. (2020)

Sha Yuan, Hanyu Zhao, Zhengxiao Du, Ming Ding, Xiao Liu, Yukuo Cen, Xu Zou, Zhilin Yang, Jie Tang: WuDaoCorpora: A super large-scale Chinese corpora for pre-training language models. AI Open, 65-68 (2021)

Yukun Zhu and Ryan Kiros and Richard S. Zemel and Ruslan Salakhutdinov and Raquel Urtasun and Antonio Torralba and Sanja Fidle: Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books. http://arxiv.org/abs/1506.06724. (2015)

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. *In Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

NLLB Team and Marta R. Costa-jussà and James Cross and Onur Çelebi and Maha Elbayad and Kenneth Heafield and Kevin Heffernan and Elahe Kalbassi and Janice Lam and Daniel Licht and Jean Maillard and Anna Sun and Skyler Wang and Guillaume Wenzek and Al Youngblood and Bapi Akula and Loic Barrault and Gabriel Mejia Gonzalez and Prangthip Hansanti and John Hoffman and Semarley Jarrett and Kaushik Ram Sadagopan and Dirk Rowe and Shannon Spruit and Chau Tran and Pierre Andrews and Necip Fazil Ayan and Shruti Bhosale and Sergey Edunov and Angela Fan and Cynthia Gao and Vedanuj Gos-wami and Francisco Guzmán and Philipp Koehn and Alexandre Mourachko and Christophe Ropers and Safiyyah Saleem and Holger Schwenk and Jeff Wang: No Language Left Be-hind: Scaling Human-Centered Machine Translation. https://arxiv.org/abs/2207.04672. (2022)

Rasley, Jeff and Rajbhandari, Samyam and Ruwase, Olatunji and He, Yuxiong: DeepSpeed: System Optimizations Enable Training Deep Learning Models with Over 100 Billion Parameters. Association for Computing Machinery, 3505-3506 (2020)

OpenAI. GPT-4 technical report. arXiv preprint arXiv:2303.08774, 2023

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

IOL Research Machine Translation Systems for WMT23 General Machine Translation Shared Task

Wenbo Zhang, Zeyu Yan, Qiaobo Deng, Jie Cai, and **Hongbao Mao**
Transn IOL Research, Wuhan, China

Abstract

This paper describes the IOL Research team's submission systems for the WMT23 general machine translation shared task. We participated in two language translation directions, including English→Chinese and Chinese→English. Our final primary submissions belong to constrained systems, which means for both translation directions we only use officially provided monolingual and bilingual data to train the translation systems. Our systems are based on Transformer architecture with pre-norm or deep-norm, which has been proven to be helpful for training deeper models. We employ methods such as back-translation, data diversification, domain fine-tuning and model ensemble to build our translation systems. An important aspect worth mentioning is our careful data cleaning process and the utilization of a substantial amount of monolingual data for data augmentation. Compared with the baseline system, our submissions have a large improvement in BLEU score.

1 Introduction

This paper describes our submissions to the WMT23 General Machine Translation shared task. We participated in two language translations: English-to-Chinese and Chinese-to-English. For both tasks, we built our system in a constrained scenario, using only official training data. Our systems are based on Transformer(Vaswani et al., 2017) architecture with pre-norm or deep-norm(Wang et al., 2022), which has been proven to be helpful for training deeper models. We used rule-based methods, language models, and alignment models to clean bilingual and monolingual data, and then used back-translation(Sennrich et al., 2016), data diversification(Nguyen et al., 2020), and model ensemble(Garmash and Monz, 2016) to leverage large-scale monolingual data to construct our translation systems. We also tried domain fine-tuning and found that this approach still helped in improving the BLEU(Papineni et al., 2002) scores on the WMT23 test set.

The design of the subsequent paper is as follows. We introduce the data source and processing strategy in Section2; Section 3 describes the details of our training procedure; Section 4 presents the experimental settings and results.

2 Data

2.1 Data Source

Bilingual corpus We used all provided bilingual data, including: ParaCrawl v9(Bañón et al., 2020), News Commentary v18.1, Wiki Titles v3, UN Parallel Corpus v1.0(Ziemski et al., 2016), CCMT Corpus, WikiMatrix(Schwenk et al., 2019), and Back-translated news.

English monolingual corpus The used English monolingual data including: News crawl, News discussions, Europarl v10, News Commentary, Common Crawl, Leipzig Corpora(Goldhahn et al., 2012), and English part of other bilingual data for WMT general task.

Chinese monolingual corpus The used Chinese monolingual data including: News crawl, News Commentary, Common Crawl, Leipzig Corpora, and Extended Common Crawl.

2.2 Data Preprocessing

For bilingual data we first filter out noisy sentences according to the rules, the filtering rules are as follows:

- Remove invisible characters.
- Remove sentences containing too more than 300 words or more than 1000 characters or less than 3 characters.
- Remove English sentences containing words exceeding than 40 characters.
- Remove Chinese sentences with a low rate of Chinese characters(less than 0.2).

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 187–191
December 6–7, 2023.

- Remove sentences that contain too many punctuation marks.
- Remove sentences that contain repeated substrings, which refers to a string composed of a single character that repeats more than 10 times, or two or more character that repeat more than 5 times.
- Remove sentences that contain HTML tags.
- Convert full-width characters to half-width characters, Traditional Chinese to Simplified Chinese.
- Remove duplicated sentence pairs.

Then we use fast-align(Dyer et al., 2013) to filter out sentence pairs with low alignment scores (less than 13) or low bilingual alignment ratio (less than 0.6), and use forward and reverse translation models to calculate the perplexity of sentence pairs, removing sentence pairs with high perplexity. For monolingual data we perform filtering using similar rules to bilingual data. At the same time, The KenLM(Heafield, 2011)[1] tool is used to train an n-gram language model to filter sentences with high perplexity scores (more than 10 000). The original parallel data totaled about 64 million sentences, and after cleaning, 46.06 million sentences were retained. Through data cleaning, we obtained 1.4 billion sentences Chinese monolingual data, and 1.2 billion sentences English monolingual data.

We used the Sentencepiece(Kudo and Richardson, 2018) tool to train the unigram model for subword segmentation, and vocabulary sizes for both Chinese and English were set to 36 000.

3 System Overview

We chose Transformer(Vaswani et al., 2017) as our base translation model and used both pre-norm and deep-norm(Wang et al., 2022) variants to help us train deeper models. To improve the quality of translation models, we first pre-trained the translation models from scratch on the synthesized datasets generated by back-translation, then continue training on the datasets generated by data diversification, and finally used domain data for fine-tuning. We also iteratively performed two rounds of data augmentation to improve the quality of the synthetic data. The final synthetic data is generated by the model after training on data diversification data of the first round. We only used domain fine-tuning in the final submission. This method we adopt is a commonly used method in the field of machine translation and has been proven to be effective. In the following sections, We show the specifics of how we use these methods.

3.1 Back-translation

Back-translation(Sennrich et al., 2016) is almost the most well-known data augmentation method in the field of machine translation, which can effectively utilize target monolingual data to improve translation quality, even in high resource situations. We used top-k sampling strategy to generate back-translation data with top-k=10, and used the method in section 2 to filter the generated data. To further increase the diversity of synthetic data, we also employed different back-translation models, such as the R2L model and the L2R model, and models with different structures to perform the back-translation method. Since this task is oriented to a general domain, we only use the cleaned monolingual data to generate synthetic data and do not select according to the domain. Because our systems are first pre-trained on back-translation data, unlike the original approach(Sennrich et al., 2016), the method back-translation in this paper refers to using only back-translation data and does not including the non-augmented corpora.

3.2 Data Diversification

Data diversification(Nguyen et al., 2020) is a data augmentation method by performing back-translation and forward-translation multiple times on the target-side and source-side data of the parallel corpus, respectively. Following this approach, we used different models to generate synthetic data by beam search. However, we not only use parallel data as source language for synthetic data, but also monolingual data. The ratio between monolingual and parallel data is 1:1.

3.3 Model Ensemble

Model ensemble can effectively improve the overall system performance by combining the strengths of multiple individual models. The larger the difference between multiple single models, the larger the improvement the ensemble model can receive. We mainly increase the diversity between single models by using different monolingual data, including different monolingual data in the back-translation

[1] https://github.com/kpu/kenlm

stage and different monolingual data in the data diversification stage.

3.4 Domain Fine-tuning

Although the WMT23 test set contains sentences from multiple domains and the WMT21 test set mainly consists of sentences from the news domain, we found that fine-tuning on the WMT21 test set can still improve the WMT23 test set. Therefore, we still attempted to fine-tune our model using newtest2021 as in-domain data.

4 Experiments

4.1 Experiment Settings

All of our translation models were implemented based on fairseq(Ott et al., 2019) and trained on 8 NVIDIA A100 GPUs. During training, we used the Adam(Kingma and Ba, 2014) optimizer with $\beta 1 = 0.9, \beta 2 = 0.98$, the learning rate scheduling strategy of inverse sqrt, the number of warmup step set to 4000, the maximum learning rate set to 0.0005 and FP16 to accelerate the training process .

We used a 24-encoder, 6-decoder transformer with pre-norm as baseline and the embedding size was set to 1024. It was trained only on a real parallel corpus, with a batch size set at 240,000 tokens. For the data augmentation models, we increased the dimension of the embedding size to 1536 and adjusted the number of the encoder and decoder layers, using equal encoder and decoder layers, or deep encoder layers and shallow decoder layers to increase the model parameter size to approximately 1 billion. The training process for these models used a batch size of 640,000 tokens. Maintaining the diversity of different models is a useful trick for model ensembles, so we trained multiple different models by adjusting the number of layers of different models, using pre-norm or deep-norm, using different synthetic data, with or without domain fine-tuning to improve diversity. Finally, we trained 4 models from Chinese to English and 5 models from English to Chinese for model ensemble.

4.2 Results

All experiments were evaluated using the sacrebleu(Post, 2018) tool to calculate BLEU(Papineni et al., 2002) scores on the WMT21, FLoRes(Goyal et al., 2021), and NTREX-128 test sets(Federmann et al., 2022). We used beam search with beam size=5 to decode all models and converted punctuation to Chinese characters in English-to-Chinese direction. Regarding the final results we submitted, we also used regular expressions for n-gram repetition detection. For translations containing repeated substrings, we set a repetition penalty of 1.5 to retranslate the source sentences. The results of Zh→En and En→Zh are shown in Table 1 and Table 2.

Based on Table 1, we can clearly see that the use of Back-Translation and Data Diversification shows significant improvements on multiple test sets. Compared to the baseline, using both data augmentation methods achieves more than 2 BLEU improvements on each test set. More than 0.5 BLEU improvement is also achieved on each test set with the model ensemble. In the end, we achieved BLEU improvements of +4.4, +3.7 and +2.8 on the three test sets of FLoRes, NTREX-128 and WMT21 respectively. The inclusion of domain fine-tuned models can further improve the WMT 23 test set compared to the model ensemble without domain fine-tuning.

From Table 2, we can see that there is a significant improvement using Back-translation on each test set. After using Data Diversification, only further improvement is achieved on the FLoRes test set, while there is varying degree of decrease on the other two test sets. Due to the decrease in diversity caused by fine-tuning multiple models with similar synthetic data generated by Data Diversification, and Data Diversification did not lead to a consistent improvement on the English to Chinese test set, in the model ensemble stage, 4 out of 5 models were trained on only Back-translation data. Finally, on the three test sets of FLoRes, NTREX-128, and WMT21, we achieve improvements of +6.5, +5.9, and +3.6 BLEUs compared to the baseline, respectively, with the model ensemble contributing the largest improvement. Similar to the results from Chinese to English, further improvements are obtained on the WMT23 test set after adding domain fine-tuning.

5 Conclusion

In this paper, we described IOL Research's submissions to the WMT2023 General Translation shared task. We participated in the English from and to Chinese translation. Our system aims to leverage as much monolingual data as possible to improve the quality of machine translation. Experimental

System	FLoRes	NTREX-128	WMT21	WMT23
Baseline	31.4	30.4	27.6	-
+Back-translation	34.2	33.2	28.4	-
+Data Diversification	35.2	33.2	29.7	-
+Ensemble	35.8	34.1	30.4	26.4
+Fine-tuning	-	-	-	27.2

Table 1: Zh→En BLEU scores on FLoRes, NTREX-128, WMT21, and WMT23 test sets. Due to the limited number of submissions, we only report part results of WMT23.

System	FLoRes	NTREX-128	WMT21	WMT23
Baseline	41.8	33.5	31.9	-
+Back-translation	44.6	37.4	33.9	-
+Data Diversification	45.2	34.5	32.8	-
+Ensemble	48.3	39.4	35.5	56.3
+Fine-tuning	-	-	-	56.9

Table 2: En→Zh BLEU scores on FLoRes, NTREX-128, WMT21, and WMT23 test sets. Due to the limited number of submissions, we only report part results of WMT23.

results show that by increasing the scale of monolingual data in the system through data augmentation and model ensemble, we have achieved substantial improvements on multiple test sets.

References

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and NoahA. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. *North American Chapter of the Association for Computational Linguistics,North American Chapter of the Association for Computational Linguistics*.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. Ntrex-128–news test references for mt evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24.

Ekaterina Garmash and Christof Monz. 2016. Ensemble learning for multi-source neural machine translation. *International Conference on Computational Linguistics,International Conference on Computational Linguistics*.

Dirk Goldhahn, Thomas Eckart, Uwe Quasthoff, et al. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *LREC*, volume 29, pages 31–43.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197.

DiederikP. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv: Learning,arXiv: Learning*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Xuan-Phi Nguyen, Shafiq Joty, Kui Wu, and Ai Ti Aw. 2020. Data diversification: A simple strategy for neural machine translation. *Advances in Neural Information Processing Systems*, 33:10018–10029.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. 2022. Deepnet: Scaling transformers to 1,000 layers. *arXiv preprint arXiv:2203.00555*.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1. 0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534.

GTCOM and DLUT's Neural Machine Translation Systems for WMT23

Hao Zong[1] **Chao Bei**[2] **Conghu Yuan**[2]
Wentao Chen[2] **Huan Liu**[2] **Degen Huang**[1]*
[1]Dalian University of Technology
[2]Global Tone Communication Technology Co., Ltd.
zonghao@mail.dlut.edu.cn
{beichao, yuanconghu, chenwentao and liuhuan}@gtcom.com.cn
huangdg@dlut.edu.cn

Abstract

This paper presents the submission by Global Tone Communication Co., Ltd. and Dalian Univeristy of Technology for the WMT23 shared general Machine Translation (MT) task at the Conference on Empirical Methods in Natural Language Processing (EMNLP). Our participation spans 8 language pairs, including English-Ukrainian, Ukrainian-English, Czech-Ukrainian, English-Hebrew, Hebrew-English, English-Czech, German-English, and Japanese-English. Our systems are designed without any specific constraints or requirements, allowing us to explore a wider range of possibilities in machine translation. We prioritize backtranslation, utilize multilingual translation models, and employ fine-tuning strategies to enhance performance. Additionally, we propose a novel data generation method that leverages human annotation to generate high-quality training data, resulting in improved system performance. Specifically, we use a combination of human-generated and machine-generated data to fine-tune our models, leading to more accurate translations. The automatic evaluation results show that our system ranks first in terms of BLEU score in Ukrainian-English, Hebrew-English, English-Hebrew, and German-English.

1 Introduction

In this study, we utilize fairseq (Ott et al., 2019) as our development tool and adopt the transformer (Vaswani et al., 2017) as the primary architecture. The main ranking index for the submitted systems is BLEU (Papineni et al., 2002), which we also employed as the evaluation metric for our translation system using sacreBLEU[1], consistent with our approach from the previous year.

For data preprocessing, we apply punctuation normalization, tokenization, and Byte Pair Encoding (BPE)(Sennrich et al., 2015) across all languages. Additionally, we applied a truecase model for English, Ukrainian and Czech, tailored to the specific characteristics of each language. In terms of tokenization, we utilized polyglot[2] for Ukrainian and Hebrew, and Moses tokenizer.perl (Koehn et al., 2007) for English and Czech. Moreover, we incorporated knowledge-based rules and a language model to clean parallel data, monolingual data, and synthetic data.

For the multilingual translation model, we amalgamated all languages into a single model and supplemented it with an English to Russian parallel corpus to enrich the language information.

The remainder of this paper is organized as follows: Section 2 introduces the translation task and presents statistics of the dataset. Section 3 describes our baseline systems and the proposed multilingual translation model. The data selection method is elaborated in Section 4. Section 5 presents experiments conducted on all translation directions, covering data filtering, model architectures, back-translation, joint training strategies, adaptations of the multilingual model, fine-tuning, data selection, and ensemble decoding. Section 6 analyzes the results, providing insights into the efficacy of different techniques. Finally, Section 7 concludes the paper.

2 Task Description

The task at hand focuses on bilingual text translation, with the provided data detailed in Table 1, which includes both parallel and monolingual data. For the English-Ukrainian and Ukrainian-English directions, the primary sources of parallel data are ParaCrawl v9 (Bañón et al., 2020), WikiMatrix (Schwenk et al., 2019), the Tilde MODEL corpus (Rozis and Skadiņš, 2017), and OPUS (Tiedemann, 2012). For the Ukrainian-Czech direction, the main parallel data comes

*Corresponding Author
[1]https://github.com/mjpost/sacrebleu
[2]https://github.com/aboSamoor/polyglot

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 192–197
December 6–7, 2023.

language	number of sentences
en-he parallel data	26.5M
en-uk parallel data	33.8M
cs-uk parallel data	6.5M
en-ru parallel data	165M
en monolingual data	90M
uk monolingual data	14M
cs monolingual data	53M
he monolingual data	5.4M
en-uk development set	1012
en-he development set	1012
cs-uk development set	1012
en-ru development set	2002
en-cs development set	1997

Table 1: Task Description

from WikiMatrix, ELRC, and OPUS. In the case of Hebrew-English and English-Hebrew, the parallel data is primarily sourced from WikiMatrix and OPUS. For English-Czech, the data sources include Europarl V10, ParaCrawl V9, Common Crawl corpus, News Commentary v18.1, CzEng 2.0 (Kocmi et al., 2020), Tilde MODEL corpus, WikiMatrix, and OPUS. For English-Russian, the sources are ParaCrawl v9, Common Crawl corpus, News Commentary v18.1, Yandex Corpus, UN Parallel Corpus V1.0(Ziemski et al., 2016), Tilde MODEL corpus, and WikiMatrix. The monolingual data utilized includes: News Crawl (Kocmi et al., 2022) in English, Ukrainian, and Czech; Leipzig Corpora (Goldhahn et al., 2012) in Hebrew, Ukrainian, and Czech; News discussions in English; News Commentary in Czech and English; and Legal Ukrainian. We used the provided development set from newstest2019 for English-Czech, newstest2020 for English-Russian, and the FLoRes101 (NLLB Team, 2022) dataset for the remaining directions.

3 Billingual Baseline Model and Multilingual Translation Model

Bilingual Baseline Model and Multilingual Translation Model: To establish a robust baseline for comparison with our multilingual model, we employed the transformer_wmt_en_de as our Bilingual baseline model, which consists of 12 encoding and 12 decoding layers. The multilingual translation model closely mirrors the GTCOM2022 (Zong and Bei, 2022) model, but this year, the focus is on the X to X model. To achieve superior translation quality, we incorporated Russian as the primary auxiliary language due to its high similarity with Ukrainian. We trained a single multilingual model that encompasses all directions. For all languages in the multilingual model, we applied joint Byte Pair Encoding (BPE) separately.

4 Data Selection

We use source test sets to train a text classification model with RoBERTa (Liu et al., 2019). Specifically, we use the in-domain test set as positive examples, and another same mount of sentence pairs from the out-of-domain test set as negative examples. We fine-tuned RoBERTa on this labeled dataset to obtain a binary classifier, which can effectively distinguish between in-domain and out-of-domain data. We then utilized this classifier to select domain-specific training data from the general training corpus. The selected in-domain training data was used to fine-tune the multilingual neural machine translation model.

We also experimented with an alternative data selection approach based on prompt learning. We constructed a prompt template and leveraged the generative power of ChatGLM-6B (Zeng et al., 2022; Du et al., 2022) to obtain an domain classifier via p-tuning (Liu et al., 2021). The prompt template is displayed in Table 2. Specifically, we extract 1,600 sentences from development set which belong to news, social, e-commerce or conversation domain. We manually select 400 sentences from training set that do not belong to domains above or are of poor quality, considering them as other domain. We then used these 2,000 labeled examples to guide the p-tuning of ChatGLM-6B. The resulting prompt-based classifier can effectively differentiate domains of training data. We consider sentences with predicted labels of "News", "Social", "E-commerce" and "Conversation" as in-domain data, and sentences with predicted labels of "Other" as out-of-domain data.

5 Experiment

This section outlines the step-by-step experiments we conducted, with the entire workflow depicted in Figure 1.

- **Data Filtering:** The data filtering methods largely replicate those we employed last year, encompassing human rules, language models, and repeat cleaning.

Instructions	Please determine the domain to which the given sentence belongs based on the following criteria. 1. Sentence Correctness: If the sentence is incomplete, incoherent, or grammatically incorrect, label it as "Other" domain. If the sentence is complete, fluent, and grammatically correct, proceed to the next step. 2. Domain Identification: Analyze the content of the sentence to identify the possible domain it belongs to. Consider the following domains: News, Social, E-commerce, Conversation, and Other. If the sentence shows clear indications of being from a specific domain, label it accordingly, otherwise label it as "Other" domain. Please label the sentence with the appropriate domain: - If the sentence is from the News domain, label it as "News". - If the sentence is from the Social domain, label it as "Social". - If the sentence is from the E-commerce domain, label it as "E-commerce". - If the sentence is from the Conversation domain, label it as "Conversation". - If the sentence does not fit any specific domain or is incorrect, label it as "Other".
Sentence	Sunday Best: Enter 1880s New York in HBO's "The Gilded Age"
Domain	News

Table 2: Prompt Template. We construct a prompt template <Instructions><Sentence><Label> for ChatGLM-6B p-tuning. Model is asked to label the <Sentence> with the appropriate domain according to <Instructions>. For each language pair in Table 1, we extract 1600 English sentences from development set and label them with given domain. Manually select 400 sentence from the training set that do not belong to specific domain or are of poor quality, and considered them as other domain. By filling <Sentence> and <Domain> with sentences above and corresponding domain, labeled samples for p-tuning can be construct.

- **Baseline:** We constructed our baseline using the transformer big architecture, which consists of 12 encoder layers and 12 decoder layers.

- **Back-translation:** We utilized the best translation model to translate the target sentence to the source side, and cleaned synthetic data with a language model. Here, we translated each language pair included in the multilingual translation model. We mixed the cleaned back-translation data and parallel sentences and trained the multilingual translation model.

- **Joint training:** We repeated the back-translation step using the best model until no further improvement was observed.

- **Multilingual translation model:** We trained a single model for all directions, with each direction having joint BPE and a shared vocabulary. The multilingual translation model comprises 24 encoder layers and 24 decoder layers, using the transformer big architecture.

- **Fine-tuning:** We fine-tuned the multilingual translation model for each direction and bi-direction separately. For instance, we fine-tuned uk2cs on the multilingual translation model and fine-tuned uk2cs and cs2uk on the multilingual translation model for Ukrainian to Czech separately.

- **Data selection:** We use model from section Data Selection to select domain-specific training dataset and fine-tune it on the multilingual translation model.

- **Ensemble Decoding:** We employed the GMSE Algorithm (Deng et al., 2018) to select models to achieve optimal performance.

6 Result and Analysis

Table 3, Table 4 and Table 5 show the BLEU score we evaluated on development set for English to/from Ukrainian, Czech to Ukrainian, English to Czech and English to/from Hebrew respectively. As shown in the above table, back-translation is still the best data augmentation measure to improve translation quality from the data aspect. Multilingual translation model also show solid improvement in all five directions. As ChatGLM only supports Chinese and English, we only perform data selection with prompt learning in

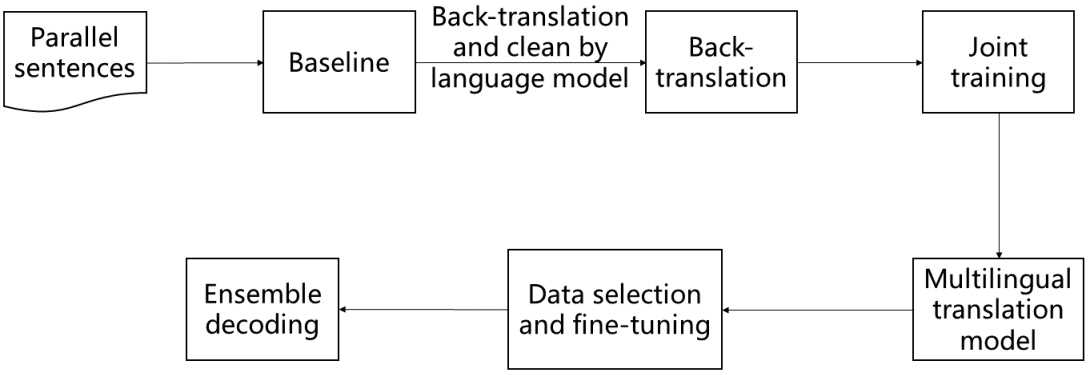


Figure 1: The work flow of GTCOM machine translation competition systems

model	en2uk	uk2en
baseline	34.11	40.99
+ back translation	34.64	41.11
multilingual translation model	34.05	40.97
+ back translation	35.01	41.96
+ bilingual fine-tuning	35.02	42.28
+ single fine-tuning	35.07	42.36
ensemble decoding	35.7	42.48

Table 3: The BLEU score between English and Ukrainian.

model	en2cs	cs2uk
baseline	28.4	23.73
+ back translation	28.61	25.45
multilingual translation model	28.29	26.05
+ back translation	28.88	27.02
+ bilingual fine-tuning	29	27.43
+ single fine-tuning	29.01	27.41
ensemble decoding	29.31	27.88

Table 4: The BLEU score of Czech to Ukrainian and English to Czech.

English-sourced language pairs. As shown in Table 6, our prompt learning strategy is still able to improve the BLEU score even after applying all other approaches. Regarding German to English and Japanese to English directions, we generate the task translations using our online system without any specific tuning.

We have noticed a significant improvement, particularly in the low-resource direction of Czech to Ukrainian, when we added Russian (which is a language closely related to Ukrainian) to the multilingual corpus.

model	en2he	he2en
baseline	34.71	45.66
+ back translation	34.8	47.06
multilingual translation model	34.52	46.74
+ back translation	35.8	46.92
+ bilingual fine-tuning	36.07	47.05
+ single fine-tuning	35.98	47.01
ensemble decoding	36.38	47.55

Table 5: The BLEU score of Czech to Ukrainian and English to Czech.

Direction	BLEU	BLEU w/o DS
en-uk	27.5	26.0
en-cs	42.3	41.1
en-he	37.2	34.6

Table 6: The final online automatic evaluation BLEU with/without prompt learning in data selection.

7 Conclusion

This paper presents GTCOM and DLUT's neural machine translation systems for the WMT23 shared general MT task. We applied three major techniques to enhance translation quality: back-translation, a multilingual translation model, and fine-tuning with data selection. By employing these techniques, we achieved significant improvements in automatic evaluation metrics, as demonstrated in Table 7.

Acknowledgments

The authors gratefully acknowledge the financial support provided by the National Key Research and Development Program of China (2020AAA0108005) and the Key Research and

Direction	BLEU
en-uk	27.5
uk-en	46.4
cs-uk	29.8
en-cs	42.3
en-he	37.2
he-en	59.2
de-en	42.2
ja-en	22.3

Table 7: The final online automatic evaluation result.

Development Program of Yunnan Province (Grant No. 202203AA080004). This work is also highly supported by 2030 Aritificial Intellegence Research Institute of Global Tone Communication Technology Co., Ltd.[3]

References

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, et al. 2020. Paracrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567.

Yongchao Deng, Shanbo Cheng, Jun Lu, Kai Song, Jingang Wang, Shenglan Wu, Liang Yao, Guchun Zhang, Haibo Zhang, Pei Zhang, et al. 2018. Alibaba's neural machine translation systems for wmt18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 368–376.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Dirk Goldhahn, Thomas Eckart, Uwe Quasthoff, et al. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *LREC*, volume 29, pages 31–43.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Martin Popel, and Ondrej Bojar. 2020. Announcing czeng 2.0 parallel corpus with over 2 gigawords. *arXiv preprint arXiv:2007.03006*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *CoRR*, abs/2110.07602.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

James Cross Onur Çelebi Maha Elbayad Kenneth Heafield Kevin Heffernan Elahe Kalbassi Janice Lam Daniel Licht Jean Maillard Anna Sun Skyler Wang Guillaume Wenzek Al Youngblood Bapi Akula Loic Barrault Gabriel Mejia Gonzalez Prangthip Hansanti John Hoffman Semarley Jarrett Kaushik Ram Sadagopan Dirk Rowe Shannon Spruit Chau Tran Pierre Andrews Necip Fazil Ayan Shruti Bhosale Sergey Edunov Angela Fan Cynthia Gao Vedanuj Goswami Francisco Guzmán Philipp Koehn Alexandre Mourachko Christophe Ropers Safiyyah Saleem Holger Schwenk Jeff Wang NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Roberts Rozis and Raivis Skadiņš. 2017. Tilde model-multilingual open data for eu languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 263–265.

[3]https://www.gtcom.com.cn

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1. 0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534.

Hao Zong and Chao Bei. 2022. GTCOM neural machine translation systems for WMT22. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 428–431, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

RoCS-MT: Robustness Challenge Set for Machine Translation

Rachel Bawden
Inria, Paris, France
rachel.bawden@inria.fr

Benoît Sagot
Inria, Paris, France
benoit.sagot@inria.fr

Abstract

RoCS-MT, a Robust Challenge Set for Machine Translation (MT), is designed to test MT systems' ability to translate user-generated content (UGC) that displays non-standard characteristics, such as spelling errors, devowelling, acronymisation, etc. RoCS-MT is composed of English comments from Reddit, selected for their non-standard nature, which have been manually normalised and professionally translated into five languages: French, German, Czech, Ukrainian and Russian. In the context of the WMT23 test suite shared task, we analyse the models submitted to the general MT task for all from-English language pairs, offering some insights into the types of problems faced by state-of-the-art MT models when dealing with non-standard UGC texts. We compare automatic metrics for MT quality, including quality estimation to see if the same conclusions can be drawn without references. In terms of robustness, we find that many of the systems struggle with non-standard variants of words (e.g. due to phonetically inspired spellings, contraction, truncations, etc.), but that this depends on the system and the amount of training data, with the best overall systems performing better across all phenomena. GPT4 is the clear frontrunner. However we caution against drawing conclusions about generalisation capacity as it and other systems could be trained on the source side of RoCS and also on similar data.

1 Introduction

As the quality of state-of-the-art machine translation (MT) systems is becoming indistinguishable in certain scenarios and domains from that of human translators (Kocmi et al., 2022), the task of tackling the translation non-standard texts is becoming an increasingly realisable aim. A considerable proportion of texts produced today are done so online in informal, unedited settings, e.g. on forums such as Twitter and Reddit, and MT is frequently to make posts accessible to a global audience. However, it has been shown that MT still struggles with user-generated content (UGC) (Gupta et al., 2023), as the type of language can differ considerably from the edited texts that have traditionally been used to train and evaluate MT models.

The RoCS-MT challenge set (Robust Challenge Set for Machine Translation) is designed to provide a test bed for the automatic translation of non-standard UGC phenomena. It contains approximately 2k sentences from the online forum Reddit that have been manually normalised and professionally translated into five languages: French, German, Czech, Ukrainian and Russian. The sentences were selected specifically for the presence of non-standard phenomena, of which we provide manual annotations (e.g. spelling errors, devowelling, capitalisations, acronymisms, etc.). Inspired by other datasets such as the French Social Media Bank (Seddah et al., 2012) and its parallel component (Rosales Núñez et al., 2019), our aim is to provide an evaluation set that is more challenging than certain previous efforts, such as the commonly used MTNT dataset (Michel and Neubig, 2018). We also make different choices from most previous efforts concerning the guidelines for normalisation and translation of the source sentences. We choose to first normalise the source sentences before translation in order to optimise the quality of the translation and to reduce the arbitrariness that may be introduced when transferring non-standard variation to the target language (e.g. on which characters to apply spelling errors, how many characters to duplicate when elongating words). For normalisation, we aim to strike a balance between normalisation as much as possible while making sure that the normalised text remains natural.[1]

In this paper, we describe the creation of the challenge set, and in the context of the WMT23 test suite shared task, we analyse the models submitted

[1] E.g. We choose to not normalise the acronym *lol* 'laughing out loud', as it is rarely/never used in its expanded form.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 198–216
December 6–7, 2023.

to the general MT shared task for the from-English shared task language pairs: English→{Czech, German, Hebrew, Japanese, Russian, Ukrainian, Chinese} (en→{cs, de, he, ja, ru, uk, zh}). Through automatic and manual analysis of system outputs, we find that many of the phenomena remain challenging for most systems (in particular those that create potential out-of-vocabulary or rare words such as phonetically inspired spellings, contractions, devowelling and truncation). However, the difficulty varies depending on the phenomenon, the particular instance (notably how frequent the non-standard word is) and the system, especially with respect to the quantity of training data. The highest performing systems overall generally do better across the board on all phenomena, whereas the weaker systems struggle in particular with certain phenomena. GPT4 has a clear lead over other systems, correctly translating even some of the most challenging examples and sometimes (although inconsistently) reproducing non-standardness in its outputs. However conclusions are limited given that the training data is unknown (as is the case of other unconstrained systems).

We make the challenge set, system outputs, evaluation code and guidelines (for the normalisation, annotation and translation) openly available for research purposes.[2]

2 Related Work

Several parallel UGC datasets exist across different language pairs. While some are extracted automatically from crawled data (Ling et al., 2013; Vicente et al., 2016; Mubarak et al., 2020), a majority are based on monolingual sentences that are then translated into the target language (Sluyter-Gäthje et al., 2018; Michel and Neubig, 2018; Rosales Núñez et al., 2019; Fujii et al., 2020; McNamee and Duh, 2022). The closest to our RoCS-MT dataset are (Michel and Neubig, 2018) and (Rosales Núñez et al., 2019), which were designed to contain challenging non-standard phenomena, whereas many of the existing datasets do not apply any such filter. Like RoCS-MT, the MTNT dataset (Michel and Neubig, 2018) contains texts from Reddit. To target non-standard language, they select sentences that have a low probability using a language model trained on standard data. In practice, and as shown by Rosales Núñez et al. (2019), the amount of non-standard language remains limited with this method. Rosales Núñez et al. (2019) base their parallel dataset on the French Social Media Bank dataset (Seddah et al., 2012), which targets non-standard language by searching for specific non-standard keywords. They show that this leads to a higher level of non-standard language, although the method is by nature more biased towards the keywords and phenomena used for data selection. An error analysis of the dataset was conducted in (Rosales Núñez et al., 2021), showing MT quality (using BLEU) for different UGC phenomena.

Despite significant effort to describe and classify UGC phenomena (Michel and Neubig, 2018; Sanguinetti et al., 2020), there is no consensus as to how texts should be normalised (and indeed translated). One extreme is to normalise all phenomena to standard forms, as is often done in lexical normalisation tasks (Han and Baldwin, 2011; van der Goot et al., 2021), but which in several cases would lead to unnatural outputs (e.g. if *lol* and *lmao*, were systematically normalised to *laughing out loud* and *laughing my ass off*). This makes translation difficult too, as the translations would also be unnatural. At the other end of the spectrum is the choice to not normalise source texts and in addition to attempt to translate the phenomena into the target language, with the disadvantage that some phenomena are language-specific[3] and others would result in arbitrary decisions being made such as to which characters to apply spelling errors. The current datasets targeting particularly non-standard phenomena choose to at least in part transfer some phenomena to the target language, whereas we adopt a higher degree of normalisation (see Section 3.1.1 for more details), producing standard but natural-sounding translations.

3 Challenge Set Creation

3.1 Data Sourcing and Selection

The source sentences are taken from English posts on discussion platform Reddit[4] using the API.[5] We do not target a particular variety of English

[2] https://github.com/rbawden/RoCS-MT

[3] Two examples of this are French *verlan*, which consists in inverting syllables in words (e.g. *louche→chelou* 'bizarre') and English cockney rhyming slang (e.g. *loaf* meaning *head* thanks to its rhyme with the expression *loaf of bread*). However, even phenomena that do exist crosslingually do not necessarily apply to the same words (e.g. the use of digits to replace their homophones as in *2day* 'today', where the translation does not necessarily contain a homophone of a digit in the target language).

[4] www.reddit.com

[5] Using the free version of the API (December 2022).

(e.g. British, American, etc.) and even include some non-native English,[6] although we do not include code-switched texts. We get an initial pool of posts by searching for specific keywords from a manually drawn-up list as in (Sanguinetti et al., 2020), e.g. *ttyl*, *ppl*, *gr8*, *alot*. The full list is given in Appendix A. For each keyword, we crawled both Reddit-wide and 3 specific subreddits (CasualUK, MadeMeSmile and entertainment) to ensure a diversity of informal topics[7] at 6-month intervals between 2017 and 2022.

Once we had the initial pool of examples, in order to reduce the number of posts to manually review, we applied a very coarse-grained in-house 'non-standardness' classifier that we had trained on a small set of manually annotated tweets according to 4 labels (standard, mildly non-standard, moderately non-standard and very non-standard), and look at posts whose title or text was marked as anything other than 'standard'. From those posts, we manually select titles and passages from the text that contain interesting non-standard phenomena, including sentences not containing the initial keyword associated with the post. This means that although our initial search process is biased to our word list, the effect is diminished by taking additional non-standard phenomena. We automatically filter out any 18+ content (using the Reddit meta-information), and manually filter out any content that is sexually inappropriate, insulting or deals with sensitive (potentially triggering) topics such as suicide or drug addiction.

3.1.1 Sentence Splitting and Normalisation

We start by manually splitting the texts into sentences. In many cases, this corresponds to splitting on final punctuation (e.g. full stop, exclamation marks, etc.). However, the non-standard nature of the texts increases the number of cases where texts are split in places that are not marked by punctuation or where punctuation or newlines are added unexpectedly in the middle of what would ordinarily be considered a sentence.

For instance, the sequence *I went grocery shopping I'm down to my last dollars soon (...)* was split into the first sentence *I went grocery shopping* and the second sentence beginning with *I'm down to my last dollars soon*, despite the lack of a final punctuation between *shopping* and *I'm*.

The first author (a native English speaker) manually normalised each of the sentences produced by our manual sentence splitting, seeking help from people knowledgeable in the topics (e.g. video gaming) where necessary. The complete normalisation guidelines with examples can be found in the dedicated Github repository.[8] As with any guidelines for dealing with complex and evolving non-standard phenomena, the decisions made are certainly not bulletproof and are likely to evolve in future work. Our aim was to reach a compromise between (i) normalising as much as possible of the text while (ii) rendering the output natural and realistic and (iii) not over-normalising such as to remove the style of the original text. We therefore normalise words such that the normalised variant could be spontaneously and naturally used.

3.1.2 Translation

Translation of the English sentences was carried out by paid professional translators. They had access to the original posts and both the raw and normalised versions of each sentence. Translation was carried out at the sentence level (following the manual segmentation and using as the source the normalised translation), although the translators had access to surrounding linguistic context, as well as additional context and translation notes provided by the first author during the normalisation step. There were also several exchanges between the first author and the translators in order to provide additional context and to answer questions. In order to preserve author anonymity, translators did not have access to meta-information about the authors (e.g. their gender). A single translation was produced for each sentence (we left the choice of speaker gender to the translators) with the exception of Ukrainian, for which two translations were produced for sentences where the speaker gender has an impact.

The target languages were chosen to cover four of those in the WMT2023 general translation task (Czech, German, Ukrainian and Russian), as well as French, which is an important language for our own research, although we do not analyse the French portion of the data in this article.

Translation Guidelines Translators were provided with guidelines (see Appendix B). They were

[6] We do not have access to any personal information about the post authors, but we know this because some posters apologise for their level of English in the posts included.

[7] The subreddits were chosen to have topics that were informal and could have a reasonable number of posts, although in reality, the number of non-standard posts found from these specific subreddits was limited.

[8] https://github.com/rbawden/RoCS-MT

instructed to translate the normalised versions of each sentence into the target language, using standard language but best matching the intention, naturalness and familiarity level of the sentence, similar to the guidelines set out in (McNamee and Duh, 2022). The decision to use standard language was to avoid the arbitrariness associated with attempting to reproduce non-standard phenomena in translation, which would make comparisons, particularly automatic ones, more difficult (e.g. which characters to alter to reproduce a spelling error, how many characters to repeat in the case of expressive repetition, etc.). They were also instructed to respect the manual segmentation provided,[9] to respect punctuation choices made in the source where appropriate (e.g. conserving full stops) and to preserve English words in meta-linguistic discussions (i.e. where authors are writing specifically about English words). As in the normalisation guidelines, abbreviations, acronyms and simplifications were to be expanded unless the result would not make a natural sentence that could realistically be found. However, abbreviations linked to the names of places and institutions were to be kept as they were if used as such in the target language (e.g. French *OTAN* for English *NATO*). They were requested not to use MT systems to help them translate in order not to bias the translations produced.

3.2 Challenge Set Subsets

We create four subsets of the challenge set to test the impact of sentence segmentation (manual or automatic using spaCy) and of normalisation (manual or none, i.e. the original raw text):

- `manseg-raw`: Manual segmentation with original (raw) text
- `manseg-norm`: Manual segmentation with manual normalisation
- `spacyseg-raw`: spaCy segmentation with original (raw) text
- `spacyseg-norm`: spaCy segmentation with manual normalisation[10]

As shown in Section 3.3, the two different segmentation methods result in different numbers of individual sentences, and automatic segmentation with spaCy differs depending on whether the text has been normalised or not. In practice, in this article, we focus only on the `manseg-raw` and `manseg-norm` subsets, although we also release the system outputs for the `spacyset-` subset. We leave research on these other subsets (i.e. looking at the impact of sentence segmentation) to future work.

3.3 Dataset Characteristics

Some basic quantitative characteristics of the data are given Table 1.

Impact of sentence splitting While the number of sentences is fixed for the manual segmentation, spaCy segmentation is highly dependent on whether the text has been normalised or not, likely due to the tool being less well adapted to non-standard text; when applied to raw text, the resulting number of sentences is far lower than manual segmentation (1660 vs. 1922), whereas the resulting number of sentences is more similar to manual segmentation when applied to the normalised text.

Tokenisation Normalisation impacts the number of tokens in the texts, as well as the number of unique tokens. When comparing the two normalised subsets on the one hand and the two raw subsets on the other (i.e. differing only in the sentence splitting), the number of tokens differs due to the fact that automatic segmentation tends to oversplit sentences on punctuation that in the manual segmentation would remain part of a token in the preceding sentence. The number of unique tokens inevitably drops after normalising, due to the homogenisation of non-standard forms (7175 vs. 6612) for manual segmentation.

Normalisation Types We manually annotated the texts for non-standard phenomena (e.g. spelling errors, acronyms, devowelling, capitalisation, pronoun drop, etc.), with the possibility of there being several types for a single span of text. Our annotations are at the word-level, with some phenomena spanning several words (e.g. capitalisation). Table 2 provides some statistics for the annotations occurring in at least 10 sentences, and some examples are given in Examples 1-4.

(1) btw I wud prefer them rlly quick.
By the way, I would prefer them really quick.
acronym contraction devow.
capitalisation
punct_diff

[9] A segment's translation can contain several sentences but sentence boundaries cannot be overridden.

[10] The spaCy segmentation was obtained by concatenating all normalised sentences from a single text and then automatically splitting.

Subset	Seg.	Norm.	#sents.	#toks.	#toks. (unique)	Ave. sent. len.	#posts	#titles	#body
`manseg-raw`	Manual	×	1922	27971	7175	14.55	391	80	263
`manseg-norm`	Manual	✓	1922	28800	6612	14.98			
`spacyseg-raw`	spaCy	×	1660	28095	7297	16.92			
`spacyseg-norm`	spaCy	✓	1996	28881	6615	14.47			

Table 1: Basic statistics of the four subsets of the test suite. Tokens are defined as whitespace delimited character sequences. Sentences can either come from post titles or the body of the post.

Annotation	#toks	#diff toks	#sents
punct_diff	2500	136	1259
capitalisation	2122	802	1059
norm_punct	542	46	339
acronymisation	329	100	277
phonetic_distance	566	285	268
spelling_error	345	306	261
spacing	294	111	250
truncation	203	104	169
contraction	161	37	146
devowelling	137	33	122
elongation	139	96	117
pronoun_drop	114	1	110
word_drop	97	2	85
grammar	75	54	73
inflection	78	64	67
lex_choice	65	52	63
article_drop	69	1	63
scrambled	38	36	37
words_to_digits	45	18	37
word_to_symbol	26	12	22
dialectism	24	15	22
double_to_single_character	17	10	17
word_add	16	13	15
digits_to_words	16	13	14
interjection	13	8	10
surrounding_emphasis	12	11	10
word_order	11	11	10
emoticon	10	10	10

Table 2: For each annotation appearing in at least 10 sentences, the number of words, unique words (lowercased) and sentences for which it appears.

(2)
So any	idea s	on	wot	I shud	be
So any	ideas	on	what	I should	be ?
	spacing		phon._dst.	contraction	punct.

(3)
Dhat	kwik	beizh	fawks	jmmpd	
That	quick	beige	fox	jumped	
phon._dst.	phon._dst.	phon._dst.	phon._dst.	phon._dst.	

(4)
Em	HOW	DARE	YOU	SWEAR	IN
EM:	How	dare	you	swear	in
caps.	caps.	caps.	caps.	caps.	caps.
punct_diff					

FRONT	OF	MY	SUN		
front	of	my	son	?	
caps.	caps.	caps.	spelling	punct_diff	

4 Translation Systems

In this article, we evaluate the systems submitted to the general translation task at WMT2023. There are both constrained and unconstrained systems, the two settings presenting significant differences in training data that should be taken into account when comparing systems.

Constrained systems Constrained systems followed similar strategies, with many systems doing data filtering/cleaning and data augmentation, using either bilingual or multilingual models and reranking. The constrained systems submitted were AIRC (Rikters and Miwa, 2023), ANVITA, CUNI-Transformer and CUNI-DocTransformer (Popel, 2020) (we refer to these system as CUNI-Trans and CUNI-DocTrans to save space in the results tables), CUNI-GA (Jon et al., 2023), HW-TSC (Wu et al., 2023b), IOL_Research (Zhang, 2023), NAIST-NICT (Deguchi et al., 2023), Samsung_Research_Philippines (Cruz, 2023) (hereafter Samsung_RP), SKIM (Kudo et al., 2023) and UvA-LTL (Wu et al., 2023a).

Unconstrained systems As in previous years of the shared task, translations were produced from anonymised online systems, corresponding in this addition to ONLINE-{A,B,G,M,W,Y} submissions. This year, translations from GPT4 were also produced using 5 few-shot examples (GPT4-5shot).[11] Note that caution should be taken when comparing results from GPT4, given that it is very possible that source sentences from RoCS-MT are included in GPT4's training data. Two systems based on NLLB (Team et al., 2022) were also submitted in the context of the metrics shared task: NLLB_Greedy and NLLB_MBR_BLEU (hereafter NLLB_MBR), which both rely on the same model but differ by the decoding strategy, either standard (greedy) or based on the Minimum Bayes Risk strategy (Freitag et al., 2022). A number of unconstrained systems were also submitted by participants, namely Lan-BridgeMT (Wu and Hu, 2023), KYB , GTCOM (Zong, 2023), (Li et al., 2023), PROMT (Molchanov and Kovalenko, 2023), Yishu

[11]The prompt used is the sentence-level prompt from (Hendy et al., 2023), which is also shown in Appendix C.

(Min et al., 2023) and ZengHuiMT (Zeng, 2023).

5 Evaluation and Analysis

Evaluation of UGC translation is more challenging than standard text; a correct translation can either be standard or non-standard in the target language, and there may be multiple ways of being non-standard that may not all be covered by available references. In our case, we chose to produce standard reference translations (See Section 3.1.2). Any system that produces non-standard language may therefore be underestimated using reference-based metrics.

We test three different metrics (BLEU, COMET and COMET-QE) to evaluate the systems' translations of RoCS-MT, looking at how coherent they are between each other, and whether it is possible to use quality estimation to evaluate MT robustness in order to remove the need for reference translations (Section 5.1). We also look at the MT quality of each system per phenomenon by calculating COMET scores over subsets of the data. Finally, we perform a qualitative analysis, manually looking at how the different systems handle UGC phenomena, and confirming some of the trends using some simple automatic analyses (Section 5.2).

5.1 Automatic evaluation

BLEU (Papineni et al., 2002), as a surface-level metric, is intuitively not robust to variation. It is therefore likely to be particularly ill-adapted to MT robustness evaluation, since MT systems' outputs can display standard or non-standard characteristics. We choose nevertheless to test this here, calculating BLEU scores using the sacreBLEU toolkit (Post, 2018).[12] We compare BLEU to reference-based COMET (Rei et al., 2020)[13] for those language pairs for which we have a reference, and to COMET's reference-less (quality estimation) version, which we refer to as COMET-QE (Rei et al., 2022).[14] We notably aim to test whether it is possible to use COMET-QE for evaluation rather than reference-based COMET, which would remove the dependency on reference translations and make evaluation possible for a wider range of languages.

Ukrainian has two reference translations for sentences for which the speaker's gender results in different translations is ambiguous between male and female. While BLEU is designed to handle multiple references, this is not inbuilt into COMET. For these sentences, we choose to take the best COMET score of the two references. For COMET and COMET-QE, which also use the source sentence, we choose to evaluate system outputs against both the `manseg-norm` and `manseg-raw` source sentences, regardless of which set was translated by the system and take the highest score of all the source-reference combinations. This covers the case where non-standard (i.e. raw) sentences are normalised during the translation process.

We provide full results for COMET and COMET-QE in Table 3 and 4 respectively, and we include results for BLEU in Table 7 in Appendix D.

How coherent are the metrics? The trends of the three metrics are similar but not at all systematic (in terms of rankings) when evaluating translations of the normalised data (`manseg-norm`), with the same systems getting the highest scores across language pairs (amongst the best systems being ONLINE-W, ONLINE-B, GPT4). However, there are some clear inconsistencies between BLEU and the two COMET metrics when evaluating non-standard data (`manseg-raw`). For example GPT4 is ranked above other systems by COMET and COMET-QE, whereas the BLEU scores of other systems (and in particular ONLINE-W and sometimes ONLINE-B) are higher. This indicates that GPT4 outputs are more surfacically different from the reference translations, which could be a result of paraphrasing or non-standard translations rather than a reflection of MT quality, especially given the high scores by COMET.

This confirms that BLEU is poorly adapted to evaluating MT robustness and could even lead to misleading conclusions, confirming previous conclusions drawn by Rosales Núñez et al. (2021) about the inadequacy of BLEU for the evaluation of UGC MT. On the other hand, COMET-QE scores show more similar trends to COMET, suggesting that it could be possible to use it to evaluate without having to produce reference translations. We nevertheless add that COMET remains an automatic metric that does not produce perfect correlation with human judgments, more research would be necessary to stress-test the metric for MT robustness evaluation, particularly in terms of evaluating which of COMET and COMET-QE is better correlated with human judgments.

[12] `case:mixed|eff:no|tok:13a|smooth:exp|v:2.2.1`
[13] We use the default `wmt22-comet-da` model.
[14] We use the default `wmt22-cometkiwi-da`.

Which systems come out on top? The highest performing systems are the unconstrained online systems, with GPT4 getting significantly higher COMET and COMET-QE scores than other systems when translating non-standard (`raw`) text for all languages tested. Other systems that tend to produce high scores are ONLINE-W and to a lesser extent ONLINE-B. Apart from these online models, both NLLB models are the best-scoring ones, which might come from the fact that they are highly multilingual and therefore could be more robust to language variation. The constrained systems, whilst not the highest performing systems, appear to get comparable scores to at least some of the online systems.

Which systems are most robust? This question is linked to the previous question about MT quality on non-standard data. To take into account the base performance of the systems, we look at the difference in score between each system's translation of the non-standard sentences and their normalised versions (also in the previously mentioned Tables 3 and 4. While there is a general trend that the higher performing systems also also have a smaller difference in quality (i.e. they are also more robust), there are some stand-out systems. GPT4 is the system with the lowest quality difference between original and normalised sentences for all language pairs tested. The NLLB models also have a low delta between the two subset, lower than or comparable to some of the more robust online systems. Similarly to the previous question, constrained systems are not the most robust in terms of their score difference. Notably for en–cs and en–de, the score differences are amongst the highest. However, some of the systems do show peformance in the same ballpark as some of the online systems.

Automatic analysis by UGC phenomenon In order to analyse how systems handle different non-standard phenomena, we evaluate sentences by annotation types, by calculating COMET and COMET-QE scores for sentences containing at least one occurrence of a particular normalisation annotation. COMET results are given in Table 5 and we include a fuller analysis for COMET-QE results in Table 8 in Appendix E. Note that we only include annotation types that appear in at least 50 sentences, and that the 'all' column refers to the scores over all sentences and not just the ones annotated for UGC phenomena.

Scores are not directly comparable across annotation type. Performance by annotation type is consistent with previous conclusions, with GPT4 getting the highest scores across the board, and online systems and NLLB also doing well. It is striking that the systems that have higher scores in general tend to do better across the board on all annotation types, whereas the lower-scoring systems struggle with certain non-standard phenomena. They correspond in particular to phonetic distance, where a word is spelt differently according to how it is pronounced (e.g. *HEERE'Z A QWESHCHUN FER YA* 'Here's a question for you'), contractions (e.g. *wud* 'would'), devowelling (e.g. *nvr* 'never'), truncation (e.g. *intro* 'introductory') and spelling errors. These are notably phenomena that could well result in out-of-vocabulary words.

Are certain language pairs more difficult than others? It is tricky to compare across language pairs, since scores are not comparable. However, there are some indications that the en–cs set is more challenging, given the low scores across multiple annotation types for all systems other than GPT4. The fact that GPT4 has high scores for all annotation types listed shows that the lower scores of other models are not due to quality issues in the reference translations, and provides an upper bound against which other systems can be compared, thereby indicating that the systems struggled more.

5.2 Qualitative analysis

Non-standard variants of words Many of the non-standard phenomena that characterise the texts (e.g. acronyms, truncations, contractions, devowelling) represent a similar difficulty to unknown or rare tokens in MT. The treatment of these words differs according to the system used, and inevitably largely on the training data of the model. Many of the constrained systems struggle to translate such words, either copying the words into the translation or omitting them entirely. The degree to which the systems succeed in correctly translating these words appears to depend on how common it is. For example, *tho*, phonetically-inspired spelling of *though*, was translated successfully by multiple systems, although the devowelled word *tmro* 'tomorrow' proved more difficult.

Markers of expressivity It is common for UGC texts to have markers of expressivity such as capitalisation or repetition of letters. We removed these markers in our normalised versions and reference

Systems	en-cs			en-de			en-ru			en-uk		
	norm	raw	Δ	norm	raw	Δ	norm	raw	Δ	norm	raw	Δ
Unconstrained												
GPT4-5shot	0.857	0.825	*0.031*	0.869	0.837	*0.032*	0.818	0.793	*0.025*	0.858	0.838	*0.021*
ONLINE-A	0.836	0.724	*0.112*	0.858	0.771	*0.087*	0.806	0.730	*0.076*	0.830	0.741	*0.090*
ONLINE-B	0.844	0.760	*0.084*	0.867	0.815	*0.052*	0.812	0.748	*0.063*	0.856	0.787	*0.069*
ONLINE-G	0.812	0.699	*0.113*	0.847	0.763	*0.084*	0.828	0.773	*0.055*	0.853	0.803	*0.050*
ONLINE-M	0.838	0.720	*0.118*	0.847	0.714	*0.133*	0.787	0.686	*0.102*	-	-	-
ONLINE-W	0.865	0.782	*0.082*	0.892	0.809	*0.083*	0.834	0.786	*0.048*	0.862	0.819	*0.043*
ONLINE-Y	0.819	0.725	*0.095*	0.862	0.795	*0.067*	0.814	0.756	*0.058*	0.823	0.750	*0.073*
NLLB_MBR	0.837	0.792	*0.045*	0.836	0.786	*0.049*	0.799	0.755	*0.045*	0.826	0.778	*0.049*
NLLB_Greedy	0.839	0.791	*0.049*	0.837	0.783	*0.054*	0.798	0.753	*0.046*	0.827	0.775	*0.052*
Lan-BridgeMT	0.820	0.723	*0.097*	0.830	0.737	*0.094*	0.784	0.699	*0.084*	0.795	0.705	*0.090*
GTCOM_Peter	0.822	0.725	*0.098*	-	-	-	-	-	-	0.807	0.714	*0.092*
PROMT	-	-	-	-	-	-	0.780	0.685	*0.095*	-	-	-
ZengHuiMT	0.811	0.717	*0.094*	0.833	0.760	*0.073*	0.772	0.706	*0.066*	0.786	0.709	*0.077*
Unconstrained												
AIRC	-	-	-	0.779	0.669	*0.110*	-	-	-	-	-	-
CUNI-Trans	0.831	0.719	*0.112*	-	-	-	-	-	-	-	-	-
CUNI-DocTrans	0.840	0.694	*0.146*	-	-	-	-	-	-	-	-	-
CUNI-GA	0.840	0.694	*0.146*	-	-	-	-	-	-	-	-	-

Table 3: COMET scores of systems on the `manseg-norm` and `manseg-raw` subsets.

Systems	en-cs			en-de			en-he			en-ja			en-ru			en-uk			en-zh		
	norm	raw	Δ	norm	raw	Δ	norm	raw	Δ	norm	raw	Δ	norm	raw	Δ	norm	raw	Δ	norm	raw	Δ
Unconstrained																					
GPT4-5shot	0.817	0.800	*0.018*	0.822	0.805	*0.017*	0.806	0.793	*0.013*	0.846	0.838	*0.008*	0.806	0.789	*0.017*	0.809	0.797	*0.012*	0.797	0.786	*0.011*
ONLINE-A	0.807	0.724	*0.083*	0.816	0.765	*0.050*	0.807	0.737	*0.070*	0.824	0.772	*0.052*	0.807	0.750	*0.058*	0.791	0.726	*0.065*	0.786	0.725	*0.061*
ONLINE-B	0.814	0.756	*0.058*	0.821	0.793	*0.028*	0.812	0.767	*0.045*	0.848	0.822	*0.027*	0.808	0.761	*0.047*	0.805	0.759	*0.046*	0.805	0.766	*0.039*
ONLINE-G	0.791	0.705	*0.086*	0.812	0.766	*0.045*	0.786	0.720	*0.067*	0.782	0.700	*0.082*	0.821	0.784	*0.036*	0.809	0.775	*0.034*	0.765	0.704	*0.062*
ONLINE-M	0.807	0.710	*0.096*	0.810	0.724	*0.086*	-	-	-	0.798	0.711	*0.088*	0.790	0.702	*0.089*	-	-	-	0.762	0.692	*0.069*
ONLINE-W	0.822	0.765	*0.057*	0.822	0.780	*0.042*	-	-	-	0.822	0.790	*0.031*	0.819	0.786	*0.033*	0.812	0.782	*0.030*	0.802	0.767	*0.036*
ONLINE-Y	0.799	0.732	*0.067*	0.822	0.786	*0.036*	0.808	0.753	*0.056*	0.842	0.811	*0.031*	0.814	0.764	*0.050*	0.787	0.731	*0.056*	0.796	0.752	*0.044*
NLLB_MBR	0.802	0.762	*0.040*	0.801	0.763	*0.038*	0.796	0.756	*0.040*	0.721	0.682	*0.039*	0.794	0.754	*0.040*	0.784	0.744	*0.040*	0.617	0.596	*0.021*
NLLB_Greedy	0.806	0.765	*0.041*	0.802	0.761	*0.041*	0.795	0.756	*0.039*	0.749	0.711	*0.038*	0.795	0.754	*0.041*	0.786	0.745	*0.041*	0.664	0.645	*0.019*
Lan-BridgeMT	0.799	0.724	*0.075*	0.805	0.741	*0.064*	0.797	0.757	*0.040*	0.827	0.774	*0.053*	0.796	0.724	*0.071*	0.769	0.696	*0.072*	0.803	0.792	*0.011*
GTCOM_Peter	0.796	0.722	*0.074*	-	-	-	0.797	0.719	*0.077*	-	-	-	-	-	-	0.774	0.704	*0.070*	-	-	-
KYB	-	-	-	-	-	-	-	-	-	0.788	0.691	*0.097*	-	-	-	-	-	-	-	-	-
PROMT	-	-	-	-	-	-	-	-	-	-	-	-	0.789	0.710	*0.079*	-	-	-	-	-	-
Yishu	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.805	0.766	*0.039*
ZengHuiMT	0.781	0.713	*0.067*	0.792	0.748	*0.045*	0.790	0.734	*0.055*	0.828	0.791	*0.037*	0.772	0.724	*0.048*	0.748	0.696	*0.052*	0.772	0.711	*0.061*
Constrained																					
AIRC	-	-	-	0.763	0.684	*0.079*	-	-	-	0.779	0.701	*0.078*	-	-	-	-	-	-	-	-	-
ANVITA	-	-	-	-	-	-	-	-	-	0.797	0.716	*0.080*	-	-	-	-	-	-	0.630	0.536	*0.094*
CUNI-Trans	0.798	0.705	*0.093*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CUNI-DocTrans	0.803	0.677	*0.126*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CUNI-GA	0.803	0.677	*0.126*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
HW-TSC	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.793	0.740	*0.054*
IOL_Research	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.770	0.696	*0.074*
NAIST-NICT	-	-	-	-	-	-	-	-	-	0.830	0.764	*0.066*	-	-	-	-	-	-	-	-	-
Samsung_RP	-	-	-	-	-	-	0.797	0.732	*0.065*	-	-	-	-	-	-	-	-	-	-	-	-
SKIM	-	-	-	-	-	-	-	-	-	0.837	0.785	*0.052*	-	-	-	-	-	-	-	-	-
UvA-LTL	-	-	-	-	-	-	0.799	0.731	*0.068*	-	-	-	-	-	-	-	-	-	-	-	-

Table 4: COMET-QE scores of systems on the `manseg-norm` and `manseg-raw` subsets.

translations for consistency. However, there are no guidelines as to how the different systems should translate them: either preserving the markers in the translation or normalising as we do in the reference. What we observe is variable behaviour depending on both the system and on the specific textual instances. Table 6 shows one of the more extreme examples of character repetition. Most of the systems fail to translate the words with character repetition and instead copy them (the case of *moooorrreeee* 'more', *poollliitte* 'polite' and *Discouuurrrse* 'Discourse'). There is greater robustness for the word *neeeeed* 'need', which is translated correctly as *brauchen* by a majority of systems, but not by AIRC, ONLINE-M and ONLINE-W, which is interesting since ONLINE-W obtains very high BLEU scores for en–de. When systems do translate the words, it tends to be the standard form that is generated (i.e. without repetition). The only example here is GPT4, which translates *moooorrreeee* as *viiieeeel* 'viel', conserving the expressivity of the source sentence. What is interesting is that this behaviour is far from consistent for GPT4, with the other non-standard words in the same example not undergoing the same treatment. This could suggest that while the system succeeded in deciphering the

System	acronymisation	article_drop	capitalisation	contraction	devowelling	elongation	grammar	inflection	lex_choice	norm_punct	phonetic_distance	pronoun_drop	punct_diff	spacing	spelling_error	truncation	word_drop	all
en-cs																		
GPT4-5shot	0.785	0.823	0.811	0.815	0.848	0.810	0.816	0.817	0.802	0.809	0.798	0.833	0.819	0.821	0.824	0.795	0.810	0.825
ONLINE-A	0.606	0.663	0.696	0.591	0.599	0.665	0.655	0.721	0.649	0.715	0.567	0.712	0.702	0.698	0.619	0.624	0.676	0.724
ONLINE-B	0.664	0.726	0.733	0.665	0.667	0.717	0.719	0.761	0.695	0.751	0.629	0.756	0.746	0.750	0.698	0.697	0.728	0.760
ONLINE-G	0.591	0.644	0.682	0.572	0.581	0.658	0.636	0.657	0.625	0.680	0.529	0.693	0.677	0.653	0.588	0.608	0.653	0.699
ONLINE-M	0.601	0.661	0.689	0.616	0.586	0.659	0.676	0.693	0.633	0.720	0.574	0.700	0.693	0.695	0.636	0.627	0.671	0.720
ONLINE-W	0.682	0.730	0.754	0.700	0.708	0.748	0.704	0.771	0.702	0.772	0.663	0.763	0.766	0.754	0.717	0.726	0.727	0.782
ONLINE-Y	0.617	0.668	0.703	0.616	0.626	0.678	0.652	0.707	0.660	0.709	0.589	0.728	0.708	0.708	0.635	0.643	0.672	0.725
NLLB_MBR	0.711	0.786	0.778	0.757	0.771	0.787	0.745	0.785	0.754	0.778	0.724	0.794	0.786	0.773	0.757	0.738	0.774	0.792
NLLB_Greedy	0.718	0.770	0.775	0.750	0.781	0.759	0.750	0.789	0.745	0.773	0.721	0.793	0.785	0.766	0.760	0.747	0.765	0.791
Lan-BridgeMT	0.618	0.670	0.701	0.612	0.594	0.674	0.650	0.693	0.651	0.709	0.573	0.730	0.703	0.693	0.632	0.636	0.674	0.723
GTCOM_Peter	0.597	0.675	0.704	0.628	0.619	0.673	0.671	0.689	0.635	0.704	0.577	0.714	0.703	0.691	0.640	0.632	0.666	0.725
ZengHuiMT	0.610	0.674	0.695	0.605	0.593	0.649	0.658	0.723	0.645	0.693	0.576	0.707	0.702	0.692	0.635	0.653	0.690	0.717
CUNI-Trans	0.605	0.655	0.683	0.591	0.612	0.663	0.635	0.670	0.644	0.719	0.563	0.699	0.693	0.677	0.622	0.632	0.669	0.719
CUNI-DocTrans	0.583	0.617	0.660	0.553	0.562	0.660	0.609	0.653	0.610	0.692	0.521	0.669	0.667	0.654	0.589	0.610	0.644	0.694
CUNI-GA	0.583	0.617	0.660	0.553	0.562	0.660	0.609	0.653	0.610	0.692	0.521	0.669	0.667	0.654	0.589	0.610	0.644	0.694
en-de																		
GPT4-5shot	0.800	0.801	0.824	0.814	0.836	0.810	0.803	0.835	0.797	0.821	0.812	0.834	0.831	0.824	0.823	0.809	0.817	0.837
ONLINE-A	0.670	0.708	0.751	0.660	0.698	0.727	0.726	0.783	0.693	0.765	0.642	0.774	0.759	0.758	0.699	0.728	0.742	0.771
ONLINE-B	0.746	0.768	0.801	0.767	0.778	0.788	0.779	0.823	0.771	0.809	0.734	0.812	0.810	0.800	0.770	0.780	0.802	0.815
ONLINE-G	0.663	0.701	0.743	0.679	0.725	0.725	0.711	0.759	0.708	0.754	0.667	0.760	0.751	0.745	0.711	0.724	0.732	0.763
ONLINE-M	0.592	0.615	0.679	0.591	0.586	0.665	0.650	0.694	0.631	0.714	0.548	0.679	0.686	0.671	0.609	0.630	0.643	0.714
ONLINE-W	0.717	0.725	0.784	0.749	0.762	0.775	0.758	0.808	0.765	0.797	0.718	0.822	0.797	0.802	0.768	0.765	0.787	0.809
ONLINE-Y	0.710	0.746	0.777	0.745	0.762	0.760	0.753	0.797	0.746	0.789	0.703	0.794	0.787	0.784	0.761	0.756	0.750	0.795
NLLB_MBR	0.702	0.749	0.776	0.751	0.756	0.763	0.724	0.786	0.729	0.768	0.718	0.780	0.778	0.767	0.744	0.735	0.755	0.786
NLLB_Greedy	0.699	0.745	0.772	0.748	0.761	0.749	0.732	0.788	0.743	0.765	0.715	0.767	0.776	0.767	0.745	0.742	0.759	0.783
Lan-BridgeMT	0.624	0.652	0.722	0.627	0.629	0.701	0.678	0.726	0.656	0.725	0.602	0.716	0.720	0.705	0.637	0.673	0.678	0.737
ZengHuiMT	0.671	0.698	0.741	0.686	0.694	0.724	0.719	0.765	0.699	0.743	0.649	0.755	0.752	0.741	0.706	0.734	0.726	0.760
AIRC	0.556	0.588	0.646	0.533	0.557	0.625	0.597	0.666	0.595	0.646	0.522	0.631	0.647	0.636	0.557	0.595	0.609	0.669
en-ru																		
GPT4-5shot	0.751	0.794	0.780	0.788	0.811	0.795	0.766	0.797	0.773	0.781	0.760	0.780	0.787	0.755	0.781	0.758	0.771	0.793
ONLINE-A	0.633	0.687	0.716	0.641	0.689	0.668	0.712	0.751	0.688	0.719	0.609	0.713	0.721	0.709	0.653	0.680	0.703	0.730
ONLINE-B	0.664	0.724	0.732	0.684	0.730	0.709	0.721	0.770	0.712	0.744	0.652	0.740	0.740	0.747	0.705	0.706	0.722	0.748
ONLINE-G	0.689	0.762	0.759	0.730	0.754	0.757	0.749	0.791	0.747	0.767	0.700	0.766	0.765	0.765	0.736	0.743	0.768	0.773
ONLINE-M	0.566	0.636	0.657	0.584	0.583	0.634	0.653	0.691	0.636	0.687	0.555	0.668	0.661	0.657	0.608	0.602	0.640	0.686
ONLINE-W	0.724	0.769	0.770	0.761	0.773	0.759	0.769	0.797	0.753	0.777	0.721	0.784	0.780	0.781	0.763	0.762	0.781	0.786
ONLINE-Y	0.668	0.738	0.742	0.708	0.723	0.714	0.730	0.781	0.727	0.741	0.668	0.747	0.747	0.751	0.724	0.706	0.730	0.756
NLLB_MBR	0.679	0.748	0.743	0.731	0.749	0.763	0.721	0.763	0.718	0.738	0.690	0.737	0.746	0.732	0.717	0.703	0.733	0.755
NLLB_Greedy	0.675	0.738	0.740	0.720	0.745	0.734	0.718	0.758	0.710	0.729	0.695	0.734	0.744	0.733	0.720	0.703	0.723	0.753
Lan-BridgeMT	0.604	0.653	0.679	0.614	0.635	0.659	0.672	0.714	0.636	0.694	0.571	0.699	0.687	0.683	0.623	0.635	0.668	0.699
PROMT	0.574	0.636	0.664	0.590	0.645	0.646	0.635	0.686	0.637	0.677	0.553	0.678	0.667	0.656	0.601	0.620	0.644	0.685
ZengHuiMT	0.623	0.694	0.691	0.615	0.669	0.647	0.687	0.730	0.663	0.683	0.600	0.700	0.696	0.690	0.658	0.657	0.688	0.706
en-uk																		
GPT4-5shot	0.804	0.838	0.828	0.834	0.834	0.818	0.817	0.836	0.804	0.831	0.809	0.834	0.832	0.835	0.826	0.811	0.819	0.838
ONLINE-A	0.626	0.688	0.725	0.635	0.665	0.679	0.705	0.744	0.680	0.736	0.608	0.728	0.724	0.721	0.656	0.676	0.721	0.741
ONLINE-B	0.701	0.744	0.770	0.723	0.748	0.742	0.766	0.790	0.750	0.788	0.679	0.774	0.779	0.788	0.734	0.744	0.764	0.787
ONLINE-G	0.718	0.788	0.791	0.771	0.773	0.779	0.773	0.807	0.767	0.797	0.736	0.797	0.795	0.795	0.766	0.775	0.808	0.803
ONLINE-W	0.763	0.783	0.802	0.805	0.805	0.789	0.807	0.831	0.808	0.818	0.757	0.805	0.814	0.817	0.787	0.793	0.812	0.819
ONLINE-Y	0.647	0.712	0.737	0.703	0.684	0.706	0.706	0.754	0.685	0.742	0.656	0.749	0.738	0.734	0.710	0.691	0.713	0.750
NLLB_MBR	0.698	0.760	0.763	0.768	0.760	0.765	0.730	0.777	0.743	0.762	0.715	0.759	0.768	0.765	0.742	0.731	0.756	0.778
NLLB_Greedy	0.692	0.740	0.761	0.741	0.758	0.724	0.732	0.767	0.730	0.756	0.703	0.768	0.766	0.762	0.742	0.735	0.757	0.775
Lan-BridgeMT	0.591	0.655	0.685	0.606	0.603	0.657	0.660	0.672	0.632	0.679	0.572	0.689	0.684	0.675	0.620	0.628	0.660	0.705
GTCOM_Peter	0.591	0.670	0.702	0.613	0.645	0.671	0.677	0.706	0.671	0.707	0.580	0.693	0.695	0.685	0.637	0.636	0.658	0.714
ZengHuiMT	0.608	0.684	0.696	0.620	0.639	0.650	0.672	0.712	0.660	0.698	0.600	0.701	0.697	0.696	0.652	0.662	0.687	0.709

Table 5: COMET scores by normalisation annotation type and averaged over all RoCS-MT sentences ('all').

non-standard English texts, there is no systematic notion of generating non-standard translations; it is possible that an expressive version of *viel* was seen far more often in the training data, thus being a probable translation in this case.

Number of copied words Bearing in mind the fact that unknown source words are often either copied or removed, we apply a simple automatic analysis to judge how often systems copy source words to diagnose a potential problem with out-of-vocabulary words. We calculate how often a sentence's translation contains words that appear in both the source and system's translations (counting occurrences for each sentences and taking the average). We calculate the difference in the average number of copied words for the`manseg-raw` and `manseg-norm` subsets (i.e. how much more are words copied when systems are translating non-standard data). Results are shown in Table 9 of Appendix F. Note that a certain number of 'copied' words is normal, as certain words are expected to be identical in source and target texts (e.g. some named entities, numbers, etc.). There are similar trends across languages: the same systems tend to have relatively fewer/more copied words in raw than norm. GPT4 has the smallest difference in number of copied words between `norm` and `raw` translations. Unsurprisingly, most other systems have more copying in `raw` than in `norm`, indicating a problem handling unknown vocabulary and confirming our previous qualitative analysis.

On a similar note, we also analysed the source-translation length ratios for both `raw` and `norm` subsets. While this could be linked to copying or omit-

	norm/raw	Text
Source	norm	“politics used to be more polite. We need a return to the discourse”
Source	raw	"politics used to be mooooorrreeee poollliitte we neeeeed a return to The Discouuurrrse"
Reference (de)	norm	„Politik früher höflicher war. Wir brauchen eine Rückkehr zum Diskurs.“
AIRC	norm	"Politik war früher höflicher. Wir brauchen eine Rückkehr zum Diskurs"
	raw	"Politiker waren früher mooooorrreeee poollliitte wir neeeed eine Rückkehr zu The Discouuurrrse"
GPT4-5shot	norm	"die Politik früher höflicher war. Wir brauchen eine Rückkehr zur Diskussion"
	raw	"die Politik früher viiieeeel höflicher war, wir brauchen eine Rückkehr zur Diskussion"
Lan-BridgeMT	norm	"die Politik früher höflicher war. Wir brauchen eine Rückkehr zum Diskurs"
	raw	"die Politik früher mooooorrreeee poollliitte war, dass wir eine Rückkehr zu The Discouuurrrse brauchten"
NLLB_MBR	norm	Politik früher höflicher war
	raw	"Politik war früher mooooorrreeeeeee poollliitte, wir brauchen eine Rückkehr zu The Discouuurrrse"
ONLINE-A	norm	"Politik früher höflicher war. Wir brauchen eine Rückkehr zum Diskurs"
	raw	"Politik früher mooooorrreeee poollliitte wir brauchten eine Rückkehr zu The Discouuurrrse"
ONLINE-B	norm	„die Politik früher höflicher war.“ Wir brauchen eine Rückkehr zum Diskurs“
	raw	„Politik früher mal mooooorrreeee poollliitte war, wir brauchen eine Rückkehr zu The Discouuurrrse“
ONLINE-G	norm	“Politik früher höflicher war. Wir brauchen eine Rückkehr zum Diskurs
	raw	"Politik früher mooooorrreeee poollliitte war, wir brauchen eine Rückkehr zur Discouurrrse"
ONLINE-M	norm	„Politik früher höflicher war. Wir brauchen eine Rückkehr zum Diskurs“
	raw	„Politik war früher mooooorrreeee poollliitte wir neeeeed a return to The Discouurrrse“
ONLINE-W	norm	"die Politik früher höflicher war. Wir brauchen eine Rückkehr zum Diskurs"
	raw	"Politik früher mooooorrreeee poollliitte we neeeeed a return to The Discouuurrrse"
ONLINE-Y	norm	„Politik früher höflicher war. Wir brauchen eine Rückkehr zum Diskurs“
	raw	„Politik früher mooooorrreeee poollliitte war, wir brauchen eine Rückkehr zu The Discouuurrrse“
ZengHuiMT	norm	"Politik früher höflicher war", war früher höflicher. Wir brauchen eine Rückkehr zum Diskurs“
	raw	"Politik früher mooooorrreeee poollliitte war, wir brauchten eine Rückkehr zu The Discouuurrrse“

Table 6: Example of character repetition linked to a mark of expressivity for en–de.

ting unknown words (as we have seen), it is more likely to be linked to overgeneration problems, linked to systems encountering text that is out-of-domain, which we occasionally observed in the system outputs. We observed that for all systems, the length ratio between `manseg-raw` translations and their source sentences was greater than those for `manseg-norm`. The effect was even greater for the texts when automatic sentence segmentation was applied (i.e. for `spacyseg-` subsets).

6 Conclusion

We have presented a new resource, RoCS-MT, a robustness challenge set for MT, designed to test MT systems on non-standard UGC. Our automatic and manual analysis show that non-standard texts are still a problem for many of the systems, including the unconstrained ones, and that certain phenomena such as phonetically inspired spellings pose a problem in particular. The comparison of COMET and COMET-QE metrics suggest that it may be possible to draw similar conclusions from automatic scoring without using references, although future work could go into more depth into analysing what is captured by the different metrics.

Limitations

The current test set is available for five from-English directions and it would be interesting to study other language directions, including those not involving English. The current version of the challenge set only contains variants for speaker gender for one of the language pairs, and we plan to add these for the other target languages in a future version.

Finally, a major limitation is one that is becoming widespread nowadays, which is that many of the systems trained and even used in research are trained on an unknown quantity of data for which the sources are unknown. Without being able to verify the fact, GPT4 and potentially some of the other systems are likely to be trained on some of the source sentences in the challenge set, and future models may even be trained on the reference translations we provide, despite it being indicated as a test set. This is a blocking factor for scientific comparison and one that goes beyond this particular resource.

Acknowledgements

This paper was funded by both authors chair positions in the PRAIRIE institute funded by the French national agency ANR as part of the “Investissements d’avenir” programme under the reference ANR-19-P3IA-0001. It was also funded by Rachel Bawden’s Emergence project, DadaNMT, funded by Sorbonne Université.

References

Jan Christian Blaise Cruz. 2023. Samsung R&D Institute Philippines at WMT 2023. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Hiroyuki Deguchi, Kenji Imamura, Yuto Nishida, Yusuke Sakai, Justin Vasselli, and Taro Watanabe. 2023. NAIST-NICT WMT'23 General MT Task Submission. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Ryo Fujii, Masato Mita, Kaori Abe, Kazuaki Hanawa, Makoto Morishita, Jun Suzuki, and Kentaro Inui. 2020. PheMT: A phenomenon-wise dataset for machine translation robustness on user-generated contents. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5929–5943, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ananya Gupta, Jae Takeuchi, and Bart Knijnenburg. 2023. On the real-world performance of machine translation: Exploring social media post-authors' perspectives. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 302–310, Toronto, Canada. Association for Computational Linguistics.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 368–378, Portland, Oregon, USA. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? A comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Josef Jon, Martin Popel, and Ondřej Bojar. 2023. CUNI-GA submission at WMT23. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Keito Kudo, Takumi Ito, Makoto Morishita, and Jun Suzuki. 2023. SKIM at WMT 2023 General Translation Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Ben Li, Yoko Matsuzaki, and Shivam Kalkar. 2023. KYB General Machine Translation Systems for WMT23. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Wang Ling, Guang Xiang, Chris Dyer, Alan Black, and Isabel Trancoso. 2013. Microblogs as parallel corpora. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 176–186, Sofia, Bulgaria. Association for Computational Linguistics.

Paul McNamee and Kevin Duh. 2022. The multilingual microblog translation corpus: Improving and evaluating translation of user-generated text. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 910–918, Marseille, France. European Language Resources Association.

Paul Michel and Graham Neubig. 2018. MTNT: A testbed for machine translation of noisy text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 543–553, Brussels, Belgium. Association for Computational Linguistics.

Luo Min, yixin tan, and Qiulin Chen. 2023. Yishu: Yishu At WMT2023 Translation Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Alexander Molchanov and Vladislav Kovalenko. 2023. PROMT Systems for WMT23 Shared General Translation Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Hamdy Mubarak, Sabit Hassan, and Ahmed Abdelali. 2020. Constructing a bilingual corpus of parallel tweets. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*, pages 14–21, Marseille, France. European Language Resources Association.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia,

Pennsylvania, USA. Association for Computational Linguistics.

Martin Popel. 2020. CUNI English-Czech and English-Polish systems in WMT20: Robust document-level training. In *Proceedings of the Fifth Conference on Machine Translation*, pages 269–273, Online. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Matīss Rikters and Makoto Miwa. 2023. AIST AIRC Submissions to the WMT23 Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

José Carlos Rosales Núñez, Djamé Seddah, and Guillaume Wisniewski. 2019. Comparison between NMT and PBSMT performance for translating noisy user-generated content. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 2–14, Turku, Finland. Linköping University Electronic Press.

José Carlos Rosales Núñez, Djamé Seddah, and Guillaume Wisniewski. 2021. Understanding the impact of UGC specificities on translation quality. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 189–198, Online. Association for Computational Linguistics.

Manuela Sanguinetti, Cristina Bosco, Lauren Cassidy, Özlem Çetinoğlu, Alessandra Teresa Cignarella, Teresa Lynn, Ines Rehbein, Josef Ruppenhofer, Djamé Seddah, and Amir Zeldes. 2020. Treebanking user-generated content: A proposal for a unified representation in Universal Dependencies. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5240–5250, Marseille, France. European Language Resources Association.

Djamé Seddah, Benoit Sagot, Marie Candito, Virginie Mouilleron, and Vanessa Combet. 2012. The French Social Media Bank: a treebank of noisy user generated content. In *Proceedings of COLING 2012*, pages 2441–2458, Mumbai, India. The COLING 2012 Organizing Committee.

Henny Sluyter-Gäthje, Pintu Lohar, Haithem Afli, and Andy Way. 2018. FooTweets: A bilingual parallel corpus of world cup tweets. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Rob van der Goot, Alan Ramponi, Arkaitz Zubiaga, Barbara Plank, Benjamin Muller, Iñaki San Vicente Roncal, Nikola Ljubešić, Özlem Çetinoğlu, Rahmad Mahendra, Talha Çolakoğlu, Timothy Baldwin, Tommaso Caselli, and Wladimir Sidorenko. 2021. MultiLexNorm: A shared task on multilingual lexical normalization. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 493–509, Online. Association for Computational Linguistics.

Iñaki San Vicente, Iñaki Alegría, Cristina España-Bonet, Pablo Gamallo, Hugo Gonçalo Oliveira, Eva Martínez Garcia, Antonio Toral, Arkaitz Zubiaga, and Nora Aranberri. 2016. TweetMT: A parallel microblog corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2936–2941, Portorož, Slovenia. European Language Resources Association (ELRA).

Di Wu, Shaomu Tan, David Stap, Ali Araabi, and Christof Monz. 2023a. UvA-MT's Participation in the WMT 2023 General Translation Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Yangjian Wu and Gang Hu. 2023. Exploring Prompt Engineering with GPT Language Models for Document-Level Machine Translation: Insights and Findings. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Zhanglin Wu, Daimeng Wei, Zongyao Li, Zhengzhe YU, Shaojun Li, Xiaoyu Chen, Hengchao Shang, Jiaxin GUO, Yuhao Xie, Lizhi Lei, Hao Yang, and Yanfei Jiang. 2023b. Treating General MT Shared Task as a Multi-Domain Adaptation Problem: HW-TSC's Submission to the WMT23 General MT Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Hui Zeng. 2023. Achieving State-of-the-Art Multilingual Translation Model with Minimal Data and Parameters. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Wenbo Zhang. 2023. IOL Research Machine Translation Systems for WMT23 General Machine Translation Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Hao Zong. 2023. Gtcom neural machine translation systems for wmt23. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

A Non-standard keywords for sourcing of posts

The full list of keywords we searched for using the Reddit API is as follows: *yyy*, *iii*, *eee*, *ppl*, *btw*, *imo*, *wtf*, *shes*, *hes*, *ima*, *shud*, *wud*, *cud*, *afaik*, *bcuz*, *hahaaa*, *dat*, *wen*, *wot*, *woz*, *bout*, *bro*, *gonna*, *lmao*, *ppl*, *smh*, *yall*, *omg*, *barley*, *fyi*, *beleive*, *seperate*, *lol*, *ttyl*, *muaha*, *mwah*, *air*, *afaik*, *fr*, *fyi*, *idk*, *ikr*, *irl*, *jk*, *nvm*, *plz*, *pls*, *cu*, *tbh*, *ur*, *wth*, *kk*, *2mo*, *2moro*, *tmrw*, *fwiw*, *nvm*, *thx*, *b4*, *ruok*, *m8*, *l8r*, *2nite*, *gr8*, *lk*, *wt*, *w/*, *peeps*, *sooo*, *verry*, *innit*, *wasnt*, *ain't*, *definately*, *yous*, *nae*, *awfull*, *freind*, *untill*, *wierd*, *aweful*, *wether* and *alot*.

The keywords were chosen as they illustrate well known non-standard phenomena, including:

- spelling errors (e.g. *wierd* 'weird', *wether* 'whether', *alot*)
- acronymisation (*nvm* 'never mind', *fyi* 'for your information)
- repetition of characters (e.g. *hahaaa*, *eee*, *sooo*)
- contractions (e.g. *cud*, *gonna*, *shud*))
- dialectisms (e.g. *ain't*, *yous*, *nae*, *innit*, *yall*)
- devowelling (e.g. *tmrw* 'tomorrow', *pls* 'please', *jk* 'joke')
- truncations, including abbrevations (e.g. *peeps* 'people', *w/* "with')
- digit phonetisation (e.g. *2nite* 'tonight', *b4* 'before', *l8r* 'later', *cu* 'see you', *ruok* 'are you ok')
- other phonetic spellings (e.g. *wot* 'what', *thx* 'thanks', *dat* 'that')
- missing whitespace (e.g. *cu* 'see you', *ruok* 'are you ok', both examples also corresponding to phonetic spellings)
- missing punctuation (e.g. *ur* (sometimes) 'you're', *wasnt* 'wasn't')
- etc.

Although the choice of keywords does create a certain bias in the types of language retrieved (especially given that several variants of some keywords are included), these keywords are used to identify posts that likely to contain other non-standard phenomena, so the final selected sentences are not restricted to those containing these keywords.

B Translation Guidelines

These guidelines are included because there are some specific constraints as to how the translations are to be carried out, and some particularities of the dataset to explain. The sentences to be translated are found in the excel spreadsheet in the column "Normalised segment". However, we also provide additional information that can help translation (see below for more information).

Origin of the text The texts to be translated are from the Reddit online forum (extracted using the API), taken from a range of different subreddits (so of different genres of text, e.g. relationship advice, advice about pets, video gaming strategy, etc.). They were selected due to their non-standard nature (spelling mistakes, abbreviations, lack of punctuation etc.).

Preprocessing of the text The texts have been manually pseudo-anonymised (usernames and names other than those representing celebrities and other well-known public figures are replaced with new names), split into "sentences" and normalised. It is the normalised versions of the sentences that are to be translated.

The sentences have been filtered to remove offensive or sensitive content (hate speech, taking drugs, suicide, etc.). However, profanities were kept as they were taken to be illustrative of the sociolect of online language. If however, you do not feel comfortable with translating something, please leave it blank and write a comment indicating that you have not translated it.

Additional context provided to help translation The text is split into short documents with one or several sentences per document. In the excel document, a sentence's document is indicated by the value in the column "Post number", and the cells are also coloured such that it is visually easier to see which sentences belong to the same document (alternating grey and white). A Reddit post is associated with a title and a text with the main content of the post. The documents can contain either the title or a subset of the text or even both. The type of text associated with each sentence is indicated in the column "Text type". Titles are marked in bold to make them visually easier to see. Although the normalised text may be sufficient to carry out the translation, we also give access to the additional information just in case:

- the title of the post
- the entire body of text associated with the post
- the raw version of the sentence (after pseudo-anonymisation and segmentation into sentences)
- some translation notes have been added to provide some context about the posts (e.g. to give an idea of what is the subject of conversation, the meaning of some expressions and abbreviations, etc. in order to make translation easier). Very occasionally there are indications about how to translate (for instance for meta-linguistic questions where people discuss particular words, it is best to keep the English words, e.g. *One word I simply can't say properly is water...* → water should be kept in English in the translation.

Constraints (important) The dataset will be used to evaluate machine translation systems on their ability to handle non-standard texts. This crucially means that: the sentence boundaries that have been defined must not be modified. It is possible to translate a sentence using several sentences if that is what is natural. However, it is not possible to merge several source sentences to produce a single translation of both (i.e. one translation per row). translators should not use machine translation systems or other computational systems to aid translation as this could bias the translations to look like translations produced by Google Translate, DeepL, ChatGPT, etc.

More specific guidelines There are multiple posts that use slang terms (e.g. gaming or general online slang such as *lol*) and it possible that the correct translation will be an English borrowing. It is fine to use an English borrowing in this case, if this is what is generally used online. The punctuation choices should be kept as much as possible, as appropriate for the target language of translation (e.g. conserving full stops, exclamation marks, quotes, etc.). As described above, there are some instances of people talking about English words, and in this case, the English words should be kept as is. Another example: *One says "Let's eat granny" making it seem like someone's going to eat their nan. However, the other example says "Lets eat, granny", implying a different meaning to the sentence.* The phrases *"Let's eat granny"* and *"Let's eat, granny"* should be kept in English. These are indicated in the translation notes.

Use of "non-standard" language:

- Any spelling mistakes that were in the raw sentence should not be reproduced in the translation (i.e. the normalised version should be used as the source sentence to translate).
- Formatting, including things like capitalisation, should (for the same reasons) follow the conventions of the normalised translation.
- Abbreviations, acronyms and simplifications (e.g. in English *wdym* = *what do you mean*, *bc* = *because*, *rly* = *really*, etc.) should be expanded, unless the result would not make a natural sentence that could realistically be found. An example of a non-natural expansion would be *lol* = *laughing out loud*, since this is not practically used.
- However, abbreviations linked to the names of places (e.g. *USA*, *UK*, *UCL* (=*University College London*) should be kept as they are if the acronym is also commonly used in the target language. In other cases, the most frequent equivalent translation should be used. (e.g. English *UN* = French *ONU*, English *NATO* = French *OTAN*).

The overall idea is that the translations should be natural and not contain the types of non-standard language that were normalised in the English versions, although they should match as best possible the style and familiarity.

Additional questions If you have any doubts or questions about the meaning of the sentences, please contact me at rachel.bawden@inria.fr to discuss things further.

C Prompt used for GPT4-5-shot

The prompt used for the GPT4-shot is the one from (Hendy et al., 2023), i.e. the following:

```
Translate this into 1. [target language]:
[shot n source]
1. [shot n reference]
Translate this into 1. [target language]:
[input]
1.
```

D BLEU scores

We provide BLEU scores for language pairs with reference translations in Table 7. The results are provided (as with the COMET scores in the main part of the paper) for the original raw subset (`manseg-raw`) and for its normalised version ((`manseg-norm`)) as well as the difference between the two scores (δ).

E COMET-QE scores by annotation type

We provide in Table 8 COMET-QE scores per annotation type for all from-English language pairs of the shared task.

F Copying analysis

Table 9 shows results for our automatic analysis of the number of source words that are found in the output translations. We calculate the number of such words, averaged over the number of sentences for each of the subsets `manseg-raw` and `manseg-norm` and we calculate the difference between the two. Positive numbers indicate that more copied words are found when systems translate the non-standard output and negative numbers indicate that more copied words are found when systems translated the normalised sentences.

Systems	en-cs			en-de			en-ru			en-uk		
	norm	raw	Δ	norm	raw	Δ	norm	raw	Δ	norm	raw	Δ
Unconstrained												
GPT4-5shot	25.9	22.5	*3.4*	46.6	40.8	*5.8*	23.4	19.6	*3.9*	27.8	25.4	*2.4*
ONLINE-A	27.1	19.3	*7.8*	49.0	38.5	*10.5*	26.1	20.4	*5.8*	31.3	25.0	*6.3*
ONLINE-B	28.4	20.9	*7.5*	47.7	40.7	*7.1*	25.7	20.6	*5.1*	39.0	29.6	*9.4*
ONLINE-G	25.0	17.4	*7.6*	46.2	35.5	*10.7*	27.9	22.5	*5.4*	29.2	25.1	*4.0*
ONLINE-M	27.8	19.1	*8.7*	44.5	29.5	*15.0*	23.6	16.5	*7.1*	-	-	
ONLINE-W	30.0	22.9	*7.2*	66.0	47.1	*18.9*	29.3	23.7	*5.6*	31.5	27.4	*4.0*
ONLINE-Y	25.6	18.8	*6.7*	48.3	39.3	*9.0*	24.7	20.1	*4.5*	30.8	25.1	*5.7*
NLLB_MBR	25.5	20.8	*4.7*	41.5	34.1	*7.4*	22.3	18.2	*4.2*	26.9	22.3	*4.6*
NLLB_Greedy	25.4	20.8	*4.6*	42.0	34.0	*8.0*	22.1	18.4	*3.7*	26.2	22.0	*4.2*
Lan-BridgeMT	26.1	18.7	*7.4*	41.3	31.2	*10.1*	22.8	17.4	*5.4*	25.8	19.9	*5.9*
GTCOM_Peter	25.3	19.2	*6.2*	-	-	-	-	-	-	26.7	21.4	*5.3*
PROMT	-	-	-	-	22.6	16.4	*6.2*	-	-			
ZengHuiMT	26.1	20.1	*6.0*	46.7	39.2	*7.5*	23.5	19.6	*3.8*	27.9	23.3	*4.6*
Constrained												
AIRC	-	-	-	35.1	24.4	*10.6*	-	-	-	-		
CUNI-Trans	27.7	19.6	*8.1*	-	-	-	-	-	-			
CUNI-DocTrans	28.9	18.0	*10.9*	-	-	-	-	-	-			
CUNI-GA	28.9	18.0	*10.9*	-	-	-	-	-	-			

Table 7: BLEU scores of systems on the `manseg-norm` and `manseg-raw` subsets.

System	acronymisation	article_drop	capitalisation	contraction	devowelling	elongation	grammar	inflection	lex_choice	norm_punct	phonetic_distance	pronoun_drop	punct_diff	spacing	spelling_error	truncation	word_drop	all
en-cs																		
GPT4-5shot	0.773	0.783	0.793	0.804	0.817	0.806	0.780	0.801	0.782	0.794	0.776	0.802	0.798	0.802	0.785	0.782	0.777	0.800
ONLINE-A	0.654	0.671	0.705	0.637	0.625	0.696	0.671	0.722	0.654	0.727	0.589	0.699	0.710	0.717	0.625	0.640	0.689	0.724
ONLINE-B	0.704	0.740	0.741	0.702	0.684	0.744	0.720	0.761	0.710	0.758	0.649	0.747	0.750	0.754	0.687	0.709	0.729	0.756
ONLINE-G	0.638	0.658	0.692	0.619	0.605	0.693	0.648	0.676	0.635	0.696	0.557	0.694	0.690	0.672	0.597	0.631	0.656	0.705
ONLINE-M	0.626	0.661	0.688	0.638	0.595	0.684	0.674	0.695	0.628	0.721	0.582	0.698	0.691	0.693	0.621	0.627	0.688	0.710
ONLINE-W	0.703	0.715	0.749	0.715	0.696	0.761	0.679	0.759	0.706	0.764	0.668	0.753	0.756	0.748	0.695	0.725	0.711	0.765
ONLINE-Y	0.677	0.686	0.719	0.673	0.659	0.719	0.675	0.718	0.669	0.726	0.622	0.725	0.725	0.717	0.645	0.673	0.686	0.732
NLLB_MBR	0.706	0.750	0.753	0.749	0.744	0.760	0.720	0.759	0.719	0.761	0.692	0.768	0.761	0.750	0.706	0.721	0.745	0.762
NLLB_Greedy	0.717	0.749	0.755	0.751	0.745	0.750	0.724	0.758	0.722	0.760	0.692	0.764	0.765	0.750	0.718	0.730	0.736	0.765
Lan-BridgeMT	0.665	0.675	0.708	0.639	0.617	0.700	0.668	0.710	0.657	0.720	0.592	0.729	0.714	0.701	0.635	0.660	0.682	0.724
GTCOM_Peter	0.643	0.660	0.708	0.653	0.631	0.709	0.672	0.686	0.642	0.715	0.599	0.713	0.710	0.702	0.630	0.644	0.680	0.722
ZengHuiMT	0.650	0.663	0.699	0.638	0.612	0.686	0.656	0.716	0.660	0.700	0.590	0.695	0.706	0.702	0.625	0.665	0.682	0.713
CUNI-Trans	0.627	0.644	0.679	0.619	0.621	0.689	0.636	0.667	0.628	0.716	0.566	0.692	0.687	0.666	0.611	0.641	0.678	0.705
CUNI-DocTrans	0.602	0.604	0.647	0.571	0.571	0.683	0.604	0.653	0.599	0.692	0.529	0.648	0.656	0.655	0.583	0.619	0.633	0.677
CUNI-GA	0.602	0.604	0.647	0.571	0.571	0.683	0.604	0.653	0.599	0.692	0.529	0.648	0.656	0.655	0.583	0.619	0.633	0.677
en-de																		
GPT4-5shot	0.795	0.790	0.800	0.797	0.819	0.804	0.795	0.813	0.785	0.806	0.788	0.798	0.804	0.803	0.796	0.783	0.784	0.805
ONLINE-A	0.720	0.725	0.755	0.692	0.725	0.740	0.735	0.769	0.719	0.769	0.668	0.758	0.762	0.759	0.698	0.737	0.749	0.765
ONLINE-B	0.770	0.767	0.789	0.760	0.772	0.784	0.774	0.797	0.766	0.798	0.729	0.783	0.794	0.787	0.755	0.770	0.786	0.793
ONLINE-G	0.725	0.730	0.758	0.713	0.742	0.756	0.730	0.763	0.731	0.767	0.689	0.761	0.763	0.758	0.712	0.738	0.746	0.766
ONLINE-M	0.658	0.649	0.702	0.646	0.624	0.710	0.676	0.717	0.658	0.737	0.596	0.697	0.707	0.687	0.635	0.671	0.674	0.724
ONLINE-W	0.738	0.725	0.769	0.748	0.753	0.761	0.745	0.789	0.752	0.782	0.715	0.790	0.777	0.776	0.741	0.749	0.759	0.780
ONLINE-Y	0.747	0.752	0.780	0.761	0.759	0.777	0.758	0.796	0.762	0.791	0.720	0.784	0.786	0.779	0.748	0.760	0.766	0.786
NLLB_MBR	0.720	0.729	0.755	0.744	0.732	0.741	0.719	0.772	0.724	0.756	0.701	0.758	0.760	0.751	0.720	0.731	0.745	0.763
NLLB_Greedy	0.721	0.721	0.752	0.744	0.736	0.739	0.730	0.769	0.733	0.763	0.702	0.748	0.759	0.751	0.718	0.731	0.750	0.761
Lan-BridgeMT	0.688	0.680	0.730	0.676	0.673	0.709	0.700	0.746	0.691	0.746	0.636	0.736	0.733	0.722	0.657	0.709	0.714	0.741
ZengHuiMT	0.710	0.700	0.737	0.708	0.700	0.728	0.718	0.763	0.708	0.746	0.662	0.732	0.745	0.739	0.699	0.734	0.716	0.748
AIRC	0.627	0.610	0.664	0.582	0.615	0.645	0.633	0.690	0.622	0.681	0.560	0.662	0.672	0.665	0.582	0.631	0.654	0.684
en-he																		
GPT4-5shot	0.772	0.791	0.789	0.794	0.816	0.796	0.771	0.798	0.780	0.790	0.760	0.784	0.791	0.792	0.777	0.780	0.770	0.793
ONLINE-A	0.664	0.693	0.726	0.658	0.654	0.697	0.698	0.719	0.692	0.729	0.621	0.725	0.728	0.727	0.649	0.683	0.695	0.737
ONLINE-B	0.715	0.719	0.762	0.712	0.741	0.756	0.732	0.767	0.733	0.766	0.682	0.763	0.765	0.756	0.711	0.734	0.726	0.767
ONLINE-G	0.650	0.655	0.711	0.658	0.662	0.694	0.677	0.702	0.666	0.713	0.634	0.707	0.708	0.701	0.643	0.668	0.677	0.720
ONLINE-Y	0.698	0.701	0.745	0.702	0.695	0.733	0.710	0.754	0.701	0.750	0.660	0.750	0.748	0.738	0.683	0.711	0.725	0.753
NLLB_MBR	0.713	0.714	0.748	0.739	0.730	0.741	0.713	0.740	0.723	0.743	0.690	0.754	0.753	0.733	0.707	0.712	0.727	0.756
NLLB_Greedy	0.712	0.724	0.750	0.738	0.740	0.741	0.718	0.752	0.725	0.746	0.691	0.753	0.753	0.746	0.713	0.730	0.718	0.756
Lan-BridgeMT	0.709	0.721	0.749	0.741	0.733	0.735	0.718	0.750	0.718	0.750	0.686	0.758	0.752	0.733	0.710	0.725	0.716	0.757
GTCOM_Peter	0.651	0.677	0.705	0.639	0.643	0.673	0.676	0.720	0.669	0.719	0.601	0.702	0.708	0.701	0.627	0.657	0.670	0.719
ZengHuiMT	0.677	0.677	0.728	0.670	0.677	0.713	0.696	0.728	0.677	0.718	0.638	0.723	0.729	0.717	0.660	0.685	0.701	0.734
Samsung_RP	0.654	0.673	0.721	0.671	0.653	0.700	0.697	0.734	0.691	0.717	0.616	0.721	0.723	0.690	0.655	0.665	0.690	0.732
UvA-LTL	0.663	0.684	0.721	0.661	0.660	0.696	0.699	0.736	0.691	0.723	0.626	0.730	0.719	0.715	0.656	0.676	0.689	0.731
en-ja																		
GPT4-5shot	0.824	0.830	0.834	0.837	0.834	0.841	0.815	0.836	0.815	0.836	0.819	0.832	0.835	0.834	0.825	0.829	0.820	0.838
ONLINE-A	0.705	0.749	0.763	0.700	0.703	0.751	0.740	0.790	0.742	0.774	0.685	0.770	0.763	0.767	0.720	0.740	0.748	0.772
ONLINE-B	0.786	0.805	0.816	0.792	0.803	0.810	0.812	0.831	0.810	0.826	0.769	0.819	0.819	0.821	0.792	0.804	0.796	0.822
ONLINE-G	0.633	0.645	0.687	0.611	0.618	0.685	0.636	0.672	0.651	0.681	0.562	0.689	0.683	0.665	0.605	0.625	0.658	0.700
ONLINE-M	0.632	0.657	0.692	0.618	0.601	0.675	0.673	0.708	0.658	0.711	0.596	0.707	0.688	0.689	0.641	0.645	0.674	0.711
ONLINE-W	0.737	0.773	0.779	0.771	0.750	0.775	0.745	0.796	0.761	0.775	0.732	0.788	0.779	0.767	0.757	0.757	0.781	0.790
ONLINE-Y	0.769	0.794	0.806	0.782	0.777	0.793	0.788	0.828	0.799	0.815	0.749	0.804	0.806	0.808	0.788	0.789	0.790	0.811
NLLB_MBR	0.612	0.643	0.670	0.639	0.638	0.676	0.644	0.659	0.619	0.641	0.611	0.678	0.668	0.644	0.619	0.610	0.649	0.682
NLLB_Greedy	0.640	0.673	0.704	0.685	0.673	0.705	0.669	0.694	0.662	0.688	0.640	0.709	0.701	0.679	0.671	0.666	0.668	0.711
Lan-BridgeMT	0.721	0.730	0.766	0.712	0.705	0.747	0.742	0.763	0.737	0.777	0.669	0.779	0.765	0.760	0.710	0.746	0.738	0.774
KYB	0.613	0.634	0.666	0.616	0.593	0.651	0.621	0.710	0.659	0.679	0.572	0.702	0.673	0.670	0.629	0.659	0.670	0.691
ZengHuiMT	0.742	0.772	0.785	0.738	0.741	0.769	0.774	0.807	0.762	0.790	0.720	0.786	0.785	0.781	0.752	0.772	0.770	0.791
AIRC	0.629	0.644	0.677	0.613	0.610	0.663	0.639	0.724	0.664	0.681	0.583	0.702	0.684	0.675	0.631	0.661	0.677	0.701
ANVITA	0.647	0.681	0.707	0.648	0.623	0.664	0.667	0.736	0.671	0.706	0.605	0.734	0.701	0.699	0.644	0.673	0.696	0.716
NAIST-NICT	0.699	0.733	0.749	0.696	0.675	0.722	0.735	0.779	0.728	0.762	0.642	0.762	0.752	0.742	0.702	0.731	0.724	0.764
SKIM	0.719	0.753	0.777	0.734	0.720	0.752	0.757	0.797	0.755	0.782	0.683	0.786	0.778	0.785	0.734	0.761	0.753	0.785
en-ru																		
GPT4-5shot	0.758	0.786	0.778	0.789	0.801	0.790	0.756	0.785	0.759	0.787	0.765	0.777	0.784	0.760	0.772	0.758	0.767	0.789
ONLINE-A	0.690	0.702	0.740	0.690	0.718	0.706	0.732	0.764	0.708	0.747	0.650	0.735	0.744	0.740	0.683	0.713	0.727	0.750
ONLINE-B	0.717	0.741	0.750	0.725	0.746	0.734	0.724	0.766	0.731	0.764	0.679	0.745	0.758	0.766	0.715	0.736	0.728	0.761
ONLINE-G	0.737	0.772	0.776	0.766	0.767	0.764	0.760	0.796	0.763	0.784	0.723	0.775	0.782	0.780	0.748	0.766	0.774	0.784
ONLINE-M	0.616	0.640	0.676	0.613	0.604	0.660	0.660	0.692	0.645	0.711	0.587	0.674	0.681	0.685	0.625	0.636	0.659	0.702
ONLINE-W	0.758	0.764	0.774	0.778	0.765	0.754	0.762	0.790	0.758	0.784	0.732	0.782	0.785	0.788	0.752	0.769	0.777	0.786
ONLINE-Y	0.695	0.750	0.751	0.727	0.725	0.730	0.735	0.775	0.746	0.761	0.683	0.758	0.759	0.762	0.726	0.723	0.741	0.764
NLLB_MBR	0.702	0.736	0.743	0.734	0.727	0.758	0.713	0.745	0.716	0.745	0.687	0.745	0.751	0.742	0.706	0.716	0.734	0.754
NLLB_Greedy	0.695	0.720	0.744	0.724	0.727	0.736	0.719	0.752	0.710	0.741	0.692	0.741	0.752	0.742	0.707	0.716	0.731	0.754
Lan-BridgeMT	0.670	0.669	0.706	0.658	0.676	0.685	0.693	0.724	0.654	0.728	0.604	0.717	0.717	0.714	0.647	0.676	0.694	0.724
PROMT	0.638	0.658	0.693	0.652	0.670	0.689	0.663	0.705	0.653	0.711	0.594	0.690	0.697	0.689	0.625	0.658	0.671	0.710
ZengHuiMT	0.680	0.706	0.716	0.661	0.692	0.684	0.705	0.743	0.679	0.711	0.630	0.707	0.719	0.714	0.676	0.697	0.701	0.724
en-uk																		
GPT4-5shot	0.777	0.794	0.789	0.806	0.807	0.794	0.771	0.790	0.772	0.795	0.770	0.792	0.795	0.796	0.784	0.777	0.774	0.797
ONLINE-A	0.656	0.686	0.713	0.655	0.677	0.682	0.688	0.726	0.672	0.727	0.619	0.716	0.715	0.713	0.651	0.674	0.698	0.726
ONLINE-B	0.707	0.731	0.749	0.716	0.732	0.747	0.738	0.759	0.726	0.761	0.672	0.747	0.756	0.763	0.709	0.724	0.733	0.759
ONLINE-G	0.726	0.756	0.768	0.759	0.760	0.765	0.746	0.776	0.750	0.778	0.714	0.769	0.772	0.773	0.740	0.759	0.763	0.775
ONLINE-W	0.748	0.760	0.771	0.779	0.774	0.765	0.755	0.791	0.770	0.783	0.721	0.776	0.780	0.785	0.744	0.763	0.765	0.782
ONLINE-Y	0.666	0.705	0.720	0.695	0.671	0.699	0.695	0.737	0.686	0.729	0.646	0.730	0.724	0.722	0.685	0.689	0.691	0.731
NLLB_MBR	0.697	0.713	0.733	0.729	0.723	0.738	0.698	0.738	0.708	0.734	0.674	0.738	0.739	0.737	0.702	0.708	0.722	0.744
NLLB_Greedy	0.693	0.700	0.734	0.717	0.714	0.719	0.702	0.748	0.714	0.736	0.675	0.751	0.740	0.741	0.707	0.717	0.732	0.745
Lan-BridgeMT	0.632	0.644	0.677	0.618	0.625	0.665	0.660	0.675	0.634	0.683	0.577	0.686	0.682	0.675	0.621	0.638	0.658	0.696
GTCOM_Peter	0.626	0.665	0.691	0.633	0.654	0.676	0.670	0.694	0.657	0.706	0.596	0.685	0.690	0.686	0.637	0.647	0.664	0.704
ZengHuiMT	0.637	0.677	0.687	0.632	0.651	0.671	0.675	0.708	0.658	0.691	0.604	0.687	0.688	0.688	0.639	0.663	0.671	0.696
en-zh																		
GPT4-5shot	0.770	0.763	0.782	0.780	0.798	0.792	0.763	0.780	0.774	0.781	0.769	0.762	0.784	0.776	0.780	0.760	0.776	0.786
ONLINE-A	0.661	0.697	0.714	0.647	0.669	0.694	0.700	0.744	0.682	0.727	0.621	0.704	0.717	0.710	0.658	0.686	0.703	0.725
ONLINE-B	0.726	0.745	0.756	0.737	0.732	0.753	0.733	0.773	0.737	0.773	0.700	0.757	0.762	0.762	0.727	0.741	0.741	0.766
ONLINE-G	0.633	0.652	0.693	0.639	0.627	0.688	0.664	0.700	0.647	0.704	0.595	0.679	0.697	0.683	0.616	0.648	0.682	0.704
ONLINE-M	0.644	0.650	0.680	0.620	0.617	0.635	0.657	0.718	0.646	0.692	0.591	0.675	0.686	0.690	0.638	0.655	0.663	0.692
ONLINE-W	0.732	0.745	0.756	0.754	0.739	0.751	0.743	0.787	0.746	0.774	0.715	0.755	0.763	0.756	0.733	0.749	0.740	0.767
ONLINE-Y	0.702	0.722	0.745	0.722	0.709	0.742	0.713	0.761	0.724	0.751	0.675	0.734	0.747	0.751	0.708	0.719	0.719	0.752
NLLB_MBR	0.522	0.531	0.587	0.569	0.555	0.565	0.551	0.551	0.503	0.556	0.521	0.585	0.578	0.551	0.521	0.510	0.555	0.596
NLLB_Greedy	0.578	0.593	0.634	0.619	0.604	0.625	0.606	0.632	0.579	0.624	0.568	0.625	0.634	0.610	0.599	0.580	0.613	0.645
Lan-BridgeMT	0.779	0.784	0.785	0.789	0.796	0.802	0.765	0.789	0.765	0.791	0.780	0.779	0.790	0.786	0.781	0.771	0.774	0.792
Yishu	0.726	0.745	0.756	0.737	0.732	0.754	0.733	0.773	0.737	0.773	0.700	0.757	0.762	0.762	0.727	0.741	0.741	0.766
ZengHuiMT	0.645	0.671	0.697	0.664	0.641	0.689	0.674	0.725	0.669	0.715	0.625	0.687	0.706	0.703	0.648	0.677	0.686	0.711
ANVITA	0.509	0.505	0.510	0.519	0.488	0.536	0.544	0.554	0.498	0.531	0.470	0.515	0.524	0.526	0.488	0.500	0.486	0.536
HW-TSC	0.677	0.705	0.729	0.685	0.680	0.717	0.704	0.744	0.709	0.744	0.651	0.722	0.734	0.732	0.685	0.707	0.699	0.740
IOL_Research	0.627	0.655	0.679	0.619	0.613	0.654	0.677	0.692	0.618	0.697	0.573	0.665	0.685	0.667	0.609	0.638	0.657	0.696

Table 8: COMET-QE scores by normalisation annotation type and averaged over all RoCS-MT sentences (‘all’).

Lang. pair	en-cs	en-de	en-he	en-ja	en-ru	en-uk	en-zh
GPT4-5shot	0.14	-0.04	-0.04	-0.06	-0.07	-0.06	-0.04
NLLB_Greedy	0.08	-0.09	0.03	-0.01	-0.03	-0.01	-0.01
NLLB_MBR	0.06	-0.04	0.03	-0.00	-0.02	-0.01	-0.02
ONLINE-W	0.62	0.31	-	0.04	0.12	0.04	0.06
ONLINE-B	0.51	0.07	0.20	0.06	0.15	0.23	0.17
ONLINE-Y	0.54	0.10	0.27	0.01	0.07	0.11	0.25
Yishu	-	-	-	-	-	-	0.17
Lan-BridgeMT	0.69	0.46	0.02	0.44	0.28	0.42	-0.04
Samsung_Research_Philippines	-	-	0.32	-	-	-	-
HW-TSC	-	-	-	-	-	-	0.18
ONLINE-G	0.77	0.26	0.01	0.74	0.11	0.11	0.54
GTCOM_Peter	0.57	-	0.63	-	-	0.34	-
ONLINE-A	0.68	0.34	0.40	0.25	0.23	0.35	0.26
UvA-LTL	-	-	0.38	-	-	-	-
SKIM	-	-	-	0.33	-	-	-
ZengHuiMT	0.72	0.32	0.39	0.27	0.33	0.42	0.26
ONLINE-M	0.64	0.74	-	0.49	0.36	-	0.46
AIRC	-	0.71	-	0.56	-	-	-
PROMT	-	-	-	-	0.46	-	-
CUNI-Trans	0.88	-	-	-	-	-	-
NAIST-NICT	-	-	-	0.61	-	-	-
IOL_Research	-	-	-	-	-	-	0.54
CUNI-DocTrans	1.53	-	-	-	-	-	-
ANVITA	-	-	-	0.62	-	-	1.90
CUNI-GA	1.53	-	-	-	-	-	-
KYB	-	-	-	1.15	-	-	-

Table 9: The difference in the number of source words present in the MT output between the `manseg-raw` and `manseg-norm` subsets, averaged across all sentences for each system. This indicates how much more (or less) source words are copied in the raw (unnormalised) sentences with respect to their normalised versions.

Multifaceted Challenge Set for Evaluating Machine Translation Performance

Xiaoyu Chen, Daimeng Wei, Zhanglin Wu, Ting Zhu, Hengchao Shang
Zongyao Li, Jiaxin Guo, Ning Xie, Lizhi Lei, Shimin Tao, Hao Yang, Yanfei Jiang
Huawei Translation Services Center, Beijing, China
{chenxiaoyu35,weidaimeng,wuzhanglin2,zhuting20,shanghengchao,
lizongyao,guojiaxin1,nicolas.xie,leilizhi,
yanghao30,jiangyanfei}@huawei.com

Abstract

Machine Translation Evaluation is critical to Machine Translation researches, as the evaluation results reflect the effectiveness of training strategies. As a result, a fair and efficient evaluation method is necessary. Many researchers have raised questions about currently available evaluation metrics from various perspectives, and propose suggestions accordingly. However, to our knowledge, few researchers has analyzed the difficulty level of source sentence and its influence on evaluation results. This paper presents HW-TSC's submission to the WMT23 MT Test Suites shared task. We propose a systematic approach for construing challenge sets from four aspects: word difficulty, length difficulty, grammar difficulty and model learning difficulty. We open-source two Multifaceted Challenge Sets for Zh→En and En→Zh. We also present results of participants in this year's General MT shared task on our test sets.

1 Introduction

Machine Translation (MT) Evaluation is an indispensable part of MT research, helping researchers verify the effectiveness of proposed training strategies and offering suggestions for future researches. However, automatic machine evaluation has raised a lot of concerns during decades of practices. One research direction is to explore the weakness of available evaluation metrics (Koehn and Monz, 2006; Callison-Burch et al., 2006; Post, 2018; Chen et al., 2022). Another direction is to analyze the soundness of test sets. For example, Freitag et al. (2020) discuss the impact of reference translationese on the evaluation results.

However, to our knowledge, few researches (Ahrenberg, 2018; Isabelle et al., 2017) has been done to discuss the influence of source sentences on the evaluation results. With the advancement of machine translations in recent years, we think that randomly sampled test sets may not be able to reflect the true gaps among models, as you can't test freshman's capability with grade-1 quiz. So we propose a strategy to collect test sentences with high-level of difficulty. The strategy considers a sentence's difficulty level from four dimensions, including word difficulty, length difficulty, grammar difficulty and model learning difficulty.

This paper presents our constructed Multifaceted Challenge Sets[1] for Zh→En and En→Zh language pairs using the strategy mentioned above. Each of the test set contains 2,000 sentences. The source sentences are from the open-sourced English Wikipedia corpus[2] while the translations are provided by our in-house translators. We report the results of participants in this year's General MT shared task on our test sets and hope to gain some insight by comparing our results with the official evaluation results.

2 Challenge Set Construction

2.1 Measuring Difficulty Level of a Test Set

We propose four indexes to measure the difficulty level of a sentence: word difficulty, length difficulty, grammar difficulty and model learning difficulty.

Word Difficulty Word difficulty is measured based on the frequency of a word appeared in the parallel training corpus. In general, the lower the frequency of a word in the training data, the more challenging for neural machine translation (NMT) to translate the word correctly.

We calculate the frequency of all words in the officially provided parallel data for the General MT shared task, and select words with frequency of more than 10 times and less than 99 times as the low-frequency word list. It should be noted that although some words fall into this frequency

[1]The test sets are open-sourced at: https://github.com/HwTsc/Multifaceted_Challenge_Set_for_MT

[2]https://dumps.Wikipediamedia.org/enWikipedia/, version 20230520 is used

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 217–223
December 6–7, 2023.

system	BLEU	chrF	COMET22	Rank$_{BLEU}$	Rank$_{chrF}$	Rank$_{COMET}$
GPT4-5SHOT	31.01	59.19	82.75	1	2	1
Lan-BridgeMT	29.81	59.45	82.16	2	1	2
ONLINE-B	29.67	57.60	80.32	3	3	3
ZengHuiMT	28.68	55.66	79.14	4	6	11
Yishu	27.64	54.95	80.14	5	8	5
ONLINE-G	27.15	57.37	80.00	6	4	6
ONLINE-A	27.08	56.25	79.99	7	5	7
ONLINE-Y	25.05	54.54	79.61	8	9	10
IOL_Research	**24.95**	**52.53**	**80.21**	**9**	**11**	**4**
HW-TSC	**24.90**	**52.56**	**79.75**	**10**	**10**	**8**
ONLINE-W	23.58	55.18	79.68	11	7	9
ONLINE-M	20.92	51.00	76.50	12	12	13
NLLB_Greedy	18.27	45.63	76.35	13	13	14
NLLB_MBR_BLEU	17.92	45.50	76.86	14	14	12
ANVITA	**16.78**	**40.85**	**75.43**	**15**	**15**	**15**

Table 1: BLEU, chrF and COMET Scores for the Zh→En translation task. Constrained systems are indicated in bold.

range, they can be divided into high-frequency subwords (e.g. newsagent = news + agent), which certainly does not meet the difficulty requirement. So we manually check the English and Chinese word lists and remove words that are consisted of high-frequency subwords. Finally we use the word lists to match the Wikipedia corpus to collect test sentences.

Length Difficulty Extremely Long and short sentences can be challenging for NMT models. In our daily practice, we find that omission and logic errors are more frequently seen in extremely long sentences. Meanwhile, due to the lack of enough context information, extremely short sentences are also error-prone.

We calculate the length (the number of English words/Chinese characters) of each sentence in the Wikipedia corpus and select 1,000 longest and shortest sentences respectively. We manually check semantics of each sentence and finally select 250 extremely long and 250 extremely short sentences as the test cases. The removed sentences include those that are incomplete, or contains obvious translationese (probably back-translation results from other languages).

Grammar Difficulty Kauchak et al. (2017) propose measuring the grammar difficulty of a sentence using the frequency of the 3rd level sentence parse tree. They employ Berkeley Parser to parse the 5.4M Wikipedia corpus and create 11 frequency bins.

Inspired by their strategy, we use Berkeley Parser to parse all sentences in the Wikipedia corpus and calculate the frequency of each 3rd level parse tree pattern. We exclude patterns that appear only once, which are highly possible to be noisy data. Then we select 1,000 sentences of which their grammar pattern has the lowest frequency as the candidate pool. Finally we manually check the semantics of each candidate and select 500 test sentences.

Model Learning Difficulty Zhao et al. (2019) observe that the translation quality is related to the entropy of the source sentence. The higher the source sentence entropy, the more likely the sentence is under-translated. They propose a formula to calculate entropy of the source sentence: Assume a word s contains K candidate translations, each of which has a probability p_k, the translation entropy for this word can be calculated by:

$$E(s) = -\sum_{k=1}^{k} p_k * \log p_k \quad (1)$$

Using this formula, we calculate the entropy of each sentence in the Wikipedia corpus and select 1,000 sentences with the highest entropy as the candidate pool. Then we manually check the semantics of each sentence and finally select 500 test cases.

system	BLEU	chrF	COMET22	Rank_{BLEU}	Rank_{chrF}	Rank_{COMET}
Yishu	48.74	45.18	86.47	1	1	2
ONLINE-B	48.72	45.17	86.47	2	2	2
ONLINE-W	45.99	42.89	86.55	3	3	1
IOL_Research	**45.28**	**41.17**	**85.29**	**4**	**4**	**6**
ONLINE-A	44.92	40.72	84.82	5	5	8
HW-TSC	**44.29**	**39.91**	**85.11**	**6**	**7**	**7**
ONLINE-Y	43.72	40.03	84.51	7	6	9
ONLINE-M	41.85	39.24	82.1	8	8	10
GPT4-5shot	41.73	38.61	85.64	9	9	4
LAN-BRIDGEMT	39.89	37.83	85.52	10	10	5
ONLINE-G	39.77	37.09	81.63	11	11	11
ZengHuiMT	35.34	31.6	81.24	12	13	12
ANVITA	**35.28**	**34.02**	**78.99**	**13**	**12**	**14**
NLLB_Greedy	30.12	27.98	79	14	14	13
NLLB_MBR_BLEU	25.84	26.02	76.62	15	15	15

Table 2: BLEU, chrF and COMET Scores for the En→Zh translation task. Constrained systems are indicated in bold.

2.2 Test Set Composition

Our Zh→En and En→Zh test sets each contains 2,000 sentences, 500 sentences per category. The source sentences are selected from the open-source Englisjh Wikipedia corpus (version 20230520)[3], using the strategy we mentioned above. The target sentences are translated by our in-house translators, without referring to any machine translation models. We recruit 10 translators whose average working experience in the translation field exceed 5 years.

3 Results and Discussions

3.1 Results on the Multifaceted Challenge Set

Table 1 and Table 2 present the Zh→En and En→Zh results, including sacreBLEU (Post, 2018), chrF (Popović, 2015), and COMET-22 (Rei et al., 2022), as well as corresponding ranks. The ranks are quite different from the official results. However, as we are unable to keep the domain distribution of our test set the same as that of the official test set, we cannot draw a conclusion of whether the ranking difference is due to different levels of source sentence difficulty or domain difference.

If the ranking difference is caused by the different difficulty levels, we can conclude that systems that perform well on average test sets may not perform as well on challenge sets. So we may need a set of test sets at different difficulty levels to comprehensively evaluate model performance. Or if the ranking difference is caused by domain issues, the top-ranked systems on the official test sets may not be so general as the task name, General MT, suggests.

We also report COMET results on each subset (see table 3 and table 4) and try to understand model performance on each dimension. According to table 3, performances of Zh→En systems vary greater under the Word and Length dimensions, as the standard deviation scores are greater than that of other dimensions and the overall result. The result indicates that incorporating low-frequent words and extremely long/short sentences into the test set may better help to significantly differ model performances. The result is similar for En→Zh translation. As shown in table 4, the standard deviation under the Word dimension is much greater than that of the overall result and other dimensions. The standard deviation under the Length category is second largest, although a little bit lower than that of the overall result.

3.2 Towards More Sound Evaluation

Automatic evaluation is still the first option for MT researchers considering its speed and cost. More reliable evaluation metrics, e.g. COMET (Rei et al., 2020), BLEURT (Sellam et al., 2020), now provide more reliable evaluation results that more align with human evaluations. Meanwhile, we believe

[3]https://dumps.Wikipediamedia.org/enWikipedia/, version 20230520 is used.

System	Vocab	Grammar	Length	Learning	overall
ANVITA	**72.32**	**77.72**	**73.52**	**78.33**	**75.43**
GPT4-5SHOT	80.84	82.65	83.81	83.67	82.75
HW-TSC	**77.66**	**80.69**	**79.02**	**81.59**	**79.75**
IOL_Research	**77.94**	**80.39**	**80.81**	**81.65**	**80.21**
Lan-BridgeMT	80.31	82.05	83.24	83.02	82.16
NLLB_Greedy	73.41	78.33	74.81	78.84	76.35
NLLB_MBR_BLEU	73.59	78.41	76.05	79.4	76.86
ONLINE-A	77.97	80.12	80.9	80.92	79.99
ONLINE-B	78.4	80.53	80.98	81.32	80.32
ONLINE-G	78.03	80.46	81.09	80.39	80
ONLINE-M	73.7	77.52	76.18	78.56	76.5
ONLINE-W	77.4	80.36	79.71	81.26	79.68
ONLINE-Y	77.21	80.43	80.07	80.7	79.61
Yishu	78.49	80.58	80.1	81.34	80.14
ZengHuiMT	77.6	79.22	80.42	79.29	79.14
Standard Deviation	2.56	1.47	2.97	1.57	2.10

Table 3: COMET22 results of Zh→En systems on each subset and on the overall challenge set, as well as the standard deviation of all systems' COMET22 scores under the category.

System	Vocab	Grammar	Length	Learning	Overall
ANVITA	**77.84**	**79.46**	**77.95**	**80.9**	**79.0**
GPT4-5SHOT	83.88	85.78	86.3	86.7	85.6
HW-TSC	**83.53**	**84.99**	**86.39**	**85.78**	**85.1**
IOL_Research	**83.63**	**85.03**	**86.87**	**85.87**	**85.3**
Lan-BridgeMT	84.27	84.94	86.76	86.18	85.5
NLLB_Greedy	74.98	79.53	81.18	80.34	79.0
NLLB_MBR_BLEU	71.62	77.33	79.99	77.59	76.6
ONLINE-A	83.25	84.33	86.22	85.68	84.8
ONLINE-B	85.94	85.58	87.21	87.37	86.5
ONLINE-G	79.43	81.18	83.41	82.68	81.6
ONLINE-M	80.32	82.57	82.05	83.74	82.1
ONLINE-W	85.52	85.99	87.66	87.18	86.6
ONLINE-Y	82.93	84.47	85.3	85.55	84.5
Yishu	85.98	85.56	87.22	87.37	86.5
ZengHuiMT	79.38	81.76	81.19	82.9	81.2
Standard Deviation	4.20	2.76	3.14	2.95	3.19

Table 4: COMET22 results of En→Zh systems on each subset and on the overall challenge set, as well as the standard deviation of all systems' COMET22 scores under the category.

there should be a more systematic approach to construct test sets. In addition to domains, we should also put difficulty level into consideration. The randomly sampled test sets represent the average difficulty level in a certain domain, which can reflect the general capability of models. However, to learn the current weakness of MT and push further researches, we need challenge sets.

4 Conclusion and Limitations

This paper presents HW-TSC's submission to the WMT23 MT Test Suites shared task. We propose increasing the test set difficulty level to better measure model performances. We propose a strategy to collect test sets with high difficulty level: word difficulty, length difficulty, grammar difficulty and model learning difficulty. We construct two multifaceted Challenge Sets for Zh→En and En→Zh directions using this strategy and report automatic evaluations of participants in this year's General MT shared task on our test sets.

However, due to time constraints, we do not perform human evaluations on the test results, which we believe will offer more insights on the performance of our challenge sets. For future researches, we will conduct direct assessment (DA) and error annotations to explore the performance of each participants on the challenge sets and compare the result with the official test sets. In addition, we will construct relatively simple test sets in the same domain, and compare the results with these challenge sets, hoping to gain more insights on the role of source sentence difficulty level.

References

Lars Ahrenberg. 2018. A challenge set for english-swedish machine translation.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of bleu in machine translation research. In *11th conference of the european chapter of the association for computational linguistics*, pages 249–256.

Xiaoyu Chen, Daimeng Wei, Hengchao Shang, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Ting Zhu, Mengli Zhu, Ning Xie, Lizhi Lei, et al. 2022. Exploring robustness of machine translation metrics: A study of twenty-two automatic metrics in the wmt22 metric task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 530–540.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. Bleu might be guilty but references are not innocent. *arXiv preprint arXiv:2004.06063*.

Pierre Isabelle, Colin Cherry, and George F. Foster. 2017. A challenge set approach to evaluating machine translation. In *Conference on Empirical Methods in Natural Language Processing*.

David Kauchak, Gondy Leroy, and Alan Hogue. 2017. Measuring text difficulty using parse-tree frequency. *Journal of the Association for Information Science and Technology*, 68(9):2088–2100.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Thibault Sellam, Amy Pu, Hyung Won Chung, Sebastian Gehrmann, Qijun Tan, Markus Freitag, Dipanjan Das, and Ankur P Parikh. 2020. Learning to evaluate translation beyond english: Bleurt submissions to the wmt metrics 2020 shared task. *arXiv preprint arXiv:2010.04297*.

Yang Zhao, Jiajun Zhang, Chengqing Zong, Zhongjun He, and Hua Wu. 2019. Addressing the under-translation problem from the entropy perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 451–458.

A Appendix

Statistical Significance for Zh→En and En→Zh

.	Lan-BridgeMT	ONLINE-B	ZenghuiMT	Yishu	ONLINE-G	ONLINE-A	ONLINE-Y	IOL_Research	HW-TSC	ONLINE-W	ONLINE-M	NLLB_Greedy	NLLB_MBR_BLEU	ANVITA
GPT4-5SHOT	1.2	1.3	2.3	3.4	3.9	3.9	6.0	6.1	6.1	7.4	10.1	12.7	13.1	14.2
Lan-BridgeMT	0.0	0.1	1.1	2.2	2.7	2.7	4.8	4.9	4.9	6.2	8.9	11.5	11.9	13.0
ONLINE-B		0.0	1.0	2.0	2.5	2.6	4.6	4.7	4.8	6.1	8.8	11.4	11.8	12.9
ZenghuiMT			0.0	1.0	1.5	1.6	3.6	3.7	3.8	5.1	7.8	10.4	10.8	11.9
Yishu				0.0	0.5	0.6	2.6	2.7	2.7	4.1	6.7	9.4	9.7	10.9
ONLINE-G					0.0	0.1	2.1	2.2	2.3	3.6	6.2	8.9	9.2	10.4
ONLINE-A						0.0	2.0	2.1	2.2	3.5	6.2	8.8	9.2	10.3
ONLINE-Y							0.0	0.1	0.2	1.5	4.1	6.8	7.1	8.3
IOL_Research								0.0	0.1	1.4	4.0	6.7	7.0	8.2
HW-TSC									0.0	1.3	4.0	6.6	7.0	8.1
ONLINE-W										0.0	2.7	5.3	5.7	6.8
ONLINE-M											0.0	2.7	3.0	4.1
NLLB_Greedy												0.0	0.3	1.5
NLLB_MBR_BLEU													0.0	1.1
ANVITA														0.0

height

Table 5: statistical significance testing of the BLEU score difference for each system pair for Zh→En. Score difference is in gray if the p-value is above 0.05

	GPT45SHOT	ONLINE-B	ONLINE-G	ONLINE-A	ZenghuiMT	ONLINE-W	Yishu	ONLINE-Y	HWTSC	IOL_Research	ONLINE-M	NLLB_Greedy	NLLB_MBR_BLEU	ANVITA
Lan-BridgeMT	0.3	1.9	2.1	3.2	3.8	4.3	4.5	4.9	6.9	6.9	8.5	13.8	14.0	18.6
GPT4-5SHOT	0.0	1.6	1.8	2.9	3.5	4.0	4.2	4.7	6.6	6.7	8.2	13.6	13.7	18.3
ONLINE-B		0.0	0.2	1.4	1.9	2.4	2.7	3.1	5.0	5.1	6.6	12.0	12.1	16.8
ONLINE-G			0.0	1.1	1.7	2.2	2.4	2.8	4.8	4.8	6.4	11.7	11.9	16.5
ONLINE-A				0.0	0.6	1.1	1.3	1.7	3.7	3.7	5.3	10.6	10.8	15.4
ZenghuiMT					0.0	0.5	0.7	1.1	3.1	3.1	4.7	10.0	10.2	14.8
ONLINE-W						0.0	0.2	0.6	2.6	2.7	4.2	9.6	9.7	14.3
Yishu							0.0	0.4	2.4	2.4	4.0	9.3	9.5	14.1
ONLINE-Y								0.0	2.0	2.0	3.5	8.9	9.0	13.7
HW-TSC									0.0	0.0	1.6	6.9	7.1	11.7
IOL_Research										0.0	1.5	6.9	7.0	11.7
ONLINE-M											0.0	5.4	5.5	10.2
NLLB_Greedy												0.0	0.1	4.8
NLLB_MBR_BLEU													0.0	4.7
ANVITA														0.0

height

Table 6: statistical significance testing of the chrF score difference for each system pair for Zh→En. Score difference is in gray if the p-value is above 0.05

	Lan-BridgeMT	ONLINE-B	IOL_Research	Yishu	ONLINE-G	ONLINE-A	HW-TSC	ONLINE-W	ONLINE-Y	ZenghuiMT	NLLB_MBR_BLEU	ONLINE-M	NLLB_Greedy	ANVITA
GPT4-5SHOT	0.6	2.4	2.5	2.6	2.8	2.8	3.0	3.1	3.1	3.6	5.9	6.3	6.4	7.3
Lan-BridgeMT	0.0	1.8	2.0	2.0	2.2	2.2	2.4	2.5	2.6	3.0	5.3	5.7	5.8	6.7
ONLINE-B		0.0	0.1	0.2	0.3	0.3	0.6	0.6	0.7	1.2	3.5	3.8	4.0	4.9
IOL_Research			0.0	0.1	0.2	0.2	0.5	0.5	0.6	1.1	3.3	3.7	3.9	4.8
Yishu				0.0	0.1	0.2	0.4	0.5	0.5	1.0	3.3	3.6	3.8	4.7
ONLINE-G					0.0	0.0	0.3	0.3	0.4	0.9	3.1	3.5	3.7	4.6
ONLINE-A						0.0	0.2	0.3	0.4	0.8	3.1	3.5	3.6	4.6
HW-TSC							0.0	0.1	0.1	0.6	2.9	3.3	3.4	4.3
ONLINE-W								0.0	0.1	0.5	2.8	3.2	3.3	4.3
ONLINE-Y									0.0	0.5	2.8	3.1	3.3	4.2
ZenghuiMT										0.0	2.3	2.6	2.8	3.7
NLLB_MBR_BLEU											0.0	0.4	0.5	1.4
ONLINE-M												0.0	0.2	1.1
NLLB_Greedy													0.0	0.9
ANVITA														0.0

height

Table 7: statistical significance testing of the COMET score difference for each system pair for Zh→En. Score difference is in gray if the p-value is above 0.05

	ONLINE-B	ONLINE-W	IOL-Research	ONLINE-A	HW-TSC	ONLINE-Y	ONLINE-M	GPT4-5shot	LAN-BRIDGEMT	ONLINE-G	ZenghuiMT	ANVITA	NLLB_Greedy	NLLB_MBR_BLEU
yishu	0.0	2.8	3.5	3.8	4.5	5.0	6.9	7.0	8.9	9.0	13.4	13.5	18.6	22.9
ONLINE-B	0.0	2.7	3.4	3.8	4.4	5.0	6.9	7.0	8.8	9.0	13.4	13.4	18.6	22.9
ONLINE-W		0.0	0.7	1.1	1.7	2.3	4.1	4.3	6.1	6.2	10.7	10.7	15.9	20.2
IOL-Research			0.0	0.4	1.0	1.6	3.4	3.6	5.4	5.5	9.9	10.0	15.2	19.4
ONLINE-A				0.0	0.6	1.2	3.1	3.2	5.0	5.2	9.6	9.6	14.8	19.1
HW-TSC					0.0	0.6	2.4	2.6	4.4	4.5	9.0	9.0	14.2	18.5
ONLINE-Y						0.0	1.9	2.0	3.8	4.0	8.4	8.4	13.6	17.9
ONLINE-M							0.0	0.1	2.0	2.1	6.5	6.6	11.7	16.0
GPT4-5shot								0.0	1.8	2.0	6.4	6.5	11.6	15.9
LAN-BRIDGEMT									0.0	0.1	4.6	4.6	9.8	14.1
ONLINE-G										0.0	4.4	4.5	9.7	13.9
ZenghuiMT											0.0	0.1	5.2	9.5
ANVITA												0.0	5.2	9.4
NLLB_Greedy													0.0	4.3
NLLB_MBR_BLEU														0.0

height

Table 8: statistical significance testing of the BLEU score difference for each system pair for En→Zh. Score difference is in gray if the p-value is above 0.05

	ONLINE-B	ONLINE-W	IOL-Research	ONLINE-A	ONLINE-Y	HW-TSC	ONLINE-M	GPT4-5shot	LAN-BRIDGEMT	ONLINE-G	ANVITA	ZenghuiMT	NLLB_Greedy	NLLB_MBR_BLEU
yishu	0.0	2.3	4.0	4.5	5.2	5.3	5.9	6.6	7.4	8.1	11.2	13.6	17.2	19.2
ONLINE-B	0.0	2.3	4.0	4.5	5.1	5.3	5.9	6.6	7.3	8.1	11.2	13.6	17.2	19.2
ONLINE-W		0.0	1.7	2.2	2.9	3.0	3.7	4.3	5.1	5.8	8.9	11.3	14.9	16.9
IOL-Research			0.0	0.5	1.1	1.3	1.9	2.6	3.3	4.1	7.2	9.6	13.2	15.2
ONLINE-A				0.0	0.7	0.8	1.5	2.1	2.9	3.6	6.7	9.1	12.7	14.7
ONLINE-Y					0.0	0.1	0.8	1.4	2.2	2.9	6.0	8.4	12.1	14.0
HW-TSC						0.0	0.7	1.3	2.1	2.8	5.9	8.3	11.9	13.9
ONLINE-M							0.0	0.6	1.4	2.2	5.2	7.6	11.3	13.2
GPT4-5shot								0.0	0.8	1.5	4.6	7.0	10.6	12.6
LAN-BRIDGEMT									0.0	0.7	3.8	6.2	9.9	11.8
ONLINE-G										0.0	3.1	5.5	9.1	11.1
ANVITA											0.0	2.4	6.0	8.0
ZenghuiMT												0.0	3.6	5.6
NLLB_Greedy													0.0	2.0
NLLB_MBR_BLEU														0.0

height

Table 9: statistical significance testing of the chrF score difference for each system pair for En→Zh. Score difference is in gray if the p-value is above 0.05

	ONLINE-B	ONLINE-W	IOL-Research	ONLINE-A	HW-TSC	ONLINE-Y	ONLINE-M	GPT4-5shot	LAN-BRIDGEMT	ONLINE-G	ZenghuiMT	ANVITA	NLLB_Greedy	NLLB_MBR_BLEU
yishu	0.1	0.1	0.9	1.0	1.3	1.4	1.7	2.0	4.5	4.9	5.3	7.6	7.6	9.9
ONLINE-B	0.0	0.0	0.8	1.0	1.2	1.4	1.7	2.0	4.4	4.8	5.2	7.5	7.5	9.8
ONLINE-W		0.0	0.8	1.0	1.2	1.4	1.7	2.0	4.4	4.8	5.2	7.5	7.5	9.8
IOL-Research			0.0	0.1	0.3	0.5	0.8	1.1	3.5	4.0	4.4	6.6	6.7	9.0
ONLINE-A				0.0	0.2	0.4	0.7	1.0	3.4	3.9	4.3	6.5	6.5	8.9
HW-TSC					0.0	0.2	0.5	0.8	3.2	3.7	4.1	6.3	6.3	8.7
ONLINE-Y						0.0	0.3	0.6	3.0	3.5	3.9	6.1	6.1	8.5
ONLINE-M							0.0	0.3	2.7	3.2	3.6	5.8	5.8	8.2
GPT4-5shot								0.0	2.4	2.9	3.3	5.5	5.5	7.9
LAN-BRIDGEMT									0.0	0.5	0.9	3.1	3.1	5.5
ONLINE-G										0.0	0.4	2.6	2.6	5.0
ZenghuiMT											0.0	2.2	2.3	4.6
ANVITA												0.0	0.0	2.4
NLLB_Greedy													0.0	2.4
NLLB_MBR_BLEU														0.0

height

Table 10: statistical significance testing of the COMET score difference for each system pair for En→Zh. Score difference is in gray if the p-value is above 0.05

Linguistically Motivated Evaluation of the 2023 State-of-the-art Machine Translation: Can GPT-4 Outperform NMT?

Shushen Manakhimova[1], Eleftherios Avramidis[1], Vivien Macketanz[1], Ekaterina Lapshinova-Koltunski[2], Sergei Bagdasarov[3] and Sebastian Möller[1]

[1]German Research Center for Artificial Intelligence (DFKI)
firstname.lastname@dfki.de
[2]University of Hildesheim, lapshinovakoltun@uni-hildesheim.de
[3]Saarland University, sergeiba@lst.uni-saarland.de

Abstract

This paper offers a fine-grained analysis of the machine translation outputs in the context of the Shared Task at the 8th Conference of Machine Translation (WMT23). Building on the foundation of previous test suite efforts, our analysis includes Large Language Models and an updated test set featuring new linguistic phenomena. To our knowledge, this is the first fine-grained linguistic analysis for the GPT-4 (5-shot) translation outputs. Our evaluation spans German–English, English–German, and English–Russian language directions. Some of the phenomena with the lowest accuracies for German–English are *idioms* and *resultative predicates*. For English–German, these include *mediopassive voice*, and *noun formation(er)*. As for English–Russian, these included *idioms* and *semantic roles*. GPT-4 (5-shot) performs equally or comparably to the best systems in German–English and English–German but falls in the second significance cluster for English–Russian.

1 Introduction

Over the past few years, we have witnessed substantial advancements in Machine Translation (MT) alongside the rapid expansion of Large Language Models (LLMs). These developments have brought translation quality up to par with human capabilities. However, these seemingly perfect translations might contain fine-grained linguistic errors that go unnoticed or get overlooked entirely in automated evaluation. A more structured approach to identifying linguistic issues in the outputs involves the use of *test suites* or *challenge sets* to systematically evaluate the system's performance on specific tasks. The current study focuses on providing a fine-grained evaluation of the translation proficiency of the latest generation of Neural Machine Translation (NMT) against the latest generation of LLMs, exemplified by ChatGPT 4.5. One of the objectives is therefore to assess whether ChatGPT, as an LLM, excels NMT in managing specific linguistic phenomena. Although our focus lies on ChatGPT, we are aware that there might be other LLMS participating in the sub-task.

In this context, we are presenting the results of the test suites analyzing state-of-the-art systems in terms of numerous linguistically motivated phenomena. These test suites[1] were applied to the MT systems submitted for evaluation at the 8th Conference on Machine Translation (WMT23; Kocmi et al., 2023) across multiple language directions: German–English, English–German, and English–Russian.

This paper is structured as follows: Section 2 goes through related work, whereas Section 3 explains how the test suite was created and applied. Section 4 outlines the setup of this year's experiment, whose results are detailed in Section 5. Section 6 concludes the paper with an outlook to future research.

2 Related Work

The origins of test suites can be traced back to the early days of machine translation in the 1990s (King and Falkedal, 1990; Way, 1991; Heid and Hildenbrand, 1991). Several researchers have adopted the use of test suites to achieve their goals. For instance, Guillou and Hardmeier (2016) employed test suites to evaluate pronoun translation. Other studies (e.g. Isabelle et al., 2017; Burchardt et al., 2017) compared different MT technologies, while Avramidis et al. (2018) explored their applicability in Quality Estimation methods.

The Machine Translation test suite track has played a significant role in this context, leading to the creation of test suites focusing on specific translation-related phenomena. For example, the work by Weller-di Marco and Fraser (2022) addressed the translation of morphologically complex

[1]https://github.com/DFKI-NLP/mt-testsuite

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 224–245
December 6–7, 2023.

words from German into English. Additionally, Semenov and Bojar's research delved into document-level translation quality assessment. These test suites, however, focus on one or at most a few phenomena per test suite, including the works by Cinkova and Bojar (2018), Bojar et al. (2018), Burlot et al. (2018), Guillou et al. (2018), Rios et al. (2018), Popović (2019), Raganato et al. (2019), Rysová et al. (2019), Vojtěchová et al. (2019), Kocmi et al. (2020), Scherrer et al. (2020), Zouhar et al. (2020). Test suites, in conjunction with human evaluation, are also instrumental in assessing the quality of machine translation metrics (Freitag et al., 2021; Avramidis and Macketanz, 2022). Our approach enables a comprehensive analysis that spans over a hundred linguistic phenomena across three language pairs (Macketanz et al., 2022a). It incorporates semi-automated human evaluation, combining efficiency with in-depth analysis. Due to our participation in past shared tasks since 2018 (Macketanz et al., 2018b), we are able to analyze the development of machine translation systems over the years.

With the growing interest surrounding LLMs, researchers have been increasingly focused on evaluating ChatGPT's performance in MT. For instance, the paper by Jiao et al. (2023) concludes that ChatGPT performs competitively with commercial translation products on high-resource European languages. A comprehensive evaluation across 18 languages of GPT models versus best-performing WMT-22 systems including human evaluations by Hendy et al. (2023) supports the previous finding. Other research explores these differences in terms of the literalness of translations produced by standard NMT and ChatGPT-3 (Raunak et al., 2023). Castilho et al. (2023) have tested ChatGPT for handling context-related linguistic phenomena such as coreference, terminology, etc. to show that it performed even better than other MT engines. This current paper also places a specific focus on evaluating ChatGPT's performance compared to other systems in the shared task.

3 Method

3.1 Test suite description

This paper focuses on three language pairs: German–English, English–German, and English–Russian. The test suite is built around specific linguistic categories, further divided into more detailed linguistic phenomena. While these categories and phenomena are specific to each language pair or direction, they may overlap across different directions. Although the logic of the test suite does not follow a particular linguistic theory, the categorization is based on linguistic research, established contrastive grammars, and findings from translation studies. The test suite was designed to cover a wide range of potential translation challenges, and its categories and phenomena were internally reviewed for objectivity by linguists and professional translators.

Test set	Test sentences	Categories	Phenomena
De–En	∼5,500	14	106
En–De	∼4,785	13	110
En–Ru	∼1232	12	51

Table 1: Metadata of the language pairs in the test suite.

Table 1 provides an overview of the number of test sentences, categories, and phenomena for each language pair. Notably, our English–Russian test set has more than doubled compared to last year, from 350 sentences (Macketanz et al., 2022b) to 1232. The new categories and phenomena have been added to the English–German direction as well.

To allow the evaluation of test sentences to operate semi-automatically, we have written rules that determine translation correctness. These rules include hand-crafted regular expressions and predefined translation outputs, applied using an internal evaluation tool (Macketanz et al., 2018a). Figure 1 illustrates the workflow of the preparation and application of our test suite.

3.2 Application of the test suite

The details regarding the development and application of our test suite are available in prior publications within the test suite track.(Macketanz et al., 2018c, 2021, 2022b; Avramidis et al., 2019, 2020). In this paper, we present an overview of the complete system. As shown in Figure 1, the building of the test suite follows steps a to c. Once test sentences are input to MT systems (step d), the test suite is applied, and automatic evaluation begins. This is done using predefined rules (step e). These rules are made of regular expressions and fixed strings, indicating correct and incorrect translations based on previous MT system outputs. Regular expressions are designed to evaluate translation accuracy for specific phenomena, possibly excluding unrelated errors. Sentences are flagged with warn-

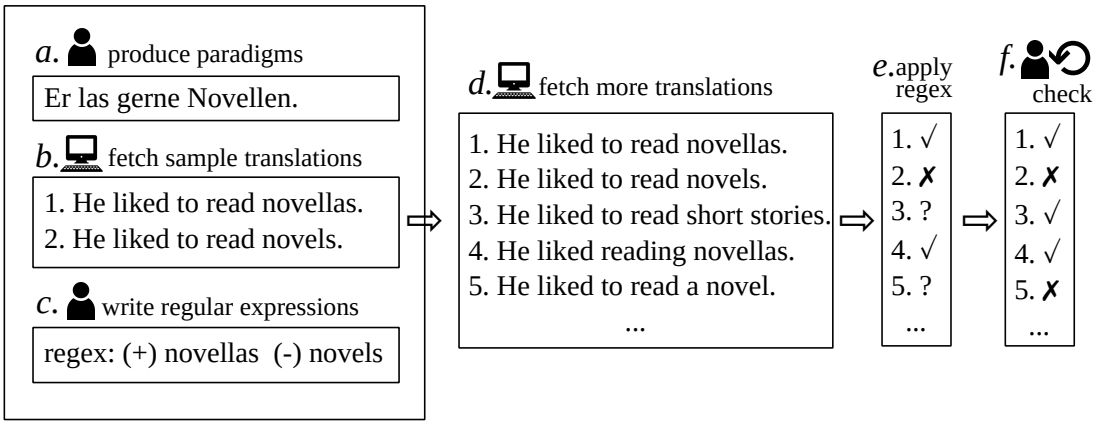


Figure 1: Example of the preparation and application of the test suite for one test sentence

ings when they cannot be automatically sorted as correct or incorrect. Human linguist annotators review and adjust the rules, while sentences with critical language errors unrelated to the phenomena are deemed incorrect.

Subsequently, the translation accuracy specific to the phenomenon is calculated by dividing the number of correctly translated test sentences for that phenomenon by the total number of test sentences for that same phenomenon:

$$\text{accuracy} = \frac{\text{correct translations}}{\text{sum of test items}}$$

Since the goal is to ensure a fair comparison among systems, only the test items that do not have any warnings are included in the calculation. If a test item has at least one unresolved warning, we exclude it from the calculation. Such an approach reduces the total number of test items, which was crucial this year, as there were many problematic outputs.

We begin by identifying the highest-scoring system in each language direction and then compare it to other systems. To do so, we confirm the significance of the comparison with a one-tailed Z-test with $\alpha = 0.95$. Systems that do not significantly differ from the top-performing system are grouped into the first performance cluster, which is indicated with boldface in the respective rows of the tables.

Average scores are computed using three distinct methods to account for variations in the number of test items within each category or phenomenon. The micro-average method aggregates the contributions of all test items to calculate average percentages. Category macro-average computes the percentages independently for each category and then averages them, treating all categories equally. Similarly, the phenomenon macro-average computes percentages independently for each phenomenon and then averages them, treating all phenomena equally.

3.3 Addition of new phenomena

This year, we added some new phenomena and made an effort to make the new test items more challenging for the systems. For instance:

- Some test items are now spanned across multiple sentences. Previously, the *coreference* category had only one sentence test items e.g., *Susan dropped the plate, and it shattered loudly*. This year, some new test items divided into two sentences had been added e.g., *The cat climbed up a tree. It was afraid*.

- There was an effort to include sentences that vary in their length, ambiguity as well syntax complexity. For example, *He was also seen wearing harem-style trousers as he tapped his feet along with his new track* as well as

- to add phenomena that require inventive approach and cultural knowledge e.g., *onomatopoeia*.

4 Experiment Setup

In this paper, we present the evaluation of 37 systems with our test suite. The systems were submitted to the *news translation task* of the Eighth Conference on Machine Translation (WMT23; Kocmi et al., 2023): 13 systems for German–English, 12 systems for English–German, and 12 systems for English–Russian.

This year is the third time that the English–German systems are being evaluated with our test suite and the second time for the English–Russian systems. Every year, manual work is involved upon receiving the system translations as there are usually a number of translation outputs that are not yet covered by the existing rules in the database (the *warnings*). At the beginning of the evaluation process this year, there were on average 10.7 % of warnings for German–English, 15.6 % for English–German, and 70.6 % for English–Russian. The English–Russian test has grown significantly since last year and in comparison with the other sets had more new items that had not been evaluated before. It was also expected that English–German would have a higher amount of warnings than German–English as there were some new categories added to the English–German test suite.

One annotator with extensive linguistic knowledge of the three languages conducted the manual evaluation of the warnings; problematic cases were discussed with several translation experts to exclude subjectivity. The manual evaluation took around three and a half weeks and involved around 55 person-hours. After the manual evaluation, there were on average 7 % of warnings left for German–English, 6.8 % for English–German, and 6.9 % for English–Russian.

As mentioned above, test sentences with at least one warning by one system were excluded from the analysis to achieve a fair comparison between the systems under inspection. As this year, we saw a lot of problematic outputs that could not be properly evaluated, this report deals with a significantly less number of test items than in the previous years. We suspect that some of these can be explained by possible models' hallucinations: a number of the MT outputs this year had some parts of the sentences repeated twice or parts of the test items were not translated at all or seemed out of place altogether. To illustrate, one unevaluated output was from the phenomenon *intransitive-perfect* "Ich bin gerannt" ("I ran" or"I was running") that in the submission of Lan-Bridge (Wu and Hu, 2023) was rendered "I'm a manager".

As a result, our analysis was conducted on 3234 (58.9 %) test sentences for German–English, 3109 (64.8 %) test sentences for English–German, and 909 (73.8 %) test sentences for English–Russian.

5 Results

All result tables can be found in the Appendix.

5.1 System comparison

For **German–English**, ChatGPT 4.5 produced micro and macro scores of 92.5 % and 91.6 % respectively, which puts ChatGPT 4.5 into the cluster of top-performing systems. The highest micro averages ranging from 95.9-93 % were achieved by the systems Online-W, Online-A, and Online-Y. In terms of the macro average, Online-W, Online-A, and Online-B demonstrated the highest scores, ranging from 91.8 % to 92.7 %. The system with the lowest performance on the micro average this year was Lan-Bridge with 81.2 %, while the system with the lowest macro average was AIRC with 74.3 %.

For the **English–German** direction, ChatGPT 4.5 leads with a micro average of 97.8 %, followed closely by Online-Y at 97.4 % and Online-B at 97.2 %. ChatGPT 4.5, on the macro average, displays the highest score 92.9 %, followed by Online-W with 92.6 % and Online-B with 92 %. The system AIRC achieved the lowest scores: 87.1 % for micro and 71 % for macro. On average, systems get micro average of 95.4 % and macro average 86.7 %.

For **English–Russian**, only Online-G and Online-W stand out with the highest scores. Online-G achieves a micro average of 86.9 % and a macro average of 86.3 %, while Online-W achieves 86.8 % and 85.5 % respectively. ChatGPT doesn't end up in the top-performing cluster and ChatGPT gets the same micro average as Online-B 81.7 %. Online-B achieves 81.3 % on macro average and outperforms ChatGPT by 3.4 %. LanguageX and Lan-Bridge as the two systems with the lowest scores achieve micro scores of 65-65.7 % and macro of 61.1 %. Several factors, such as limited training data and substantial structural differences between the languages, contribute to the translation challenges for this language pair, compared to the relatively similar English–German pair.

5.2 Category-level analysis

In **German–English**, a few models achieve 100 % in categories such as *composition*, *named entity & terminology*, and *negation*. This might be attributed to the fact that these categories have well-defined rules that the models have mastered. Categories like *ambiguity* and *false friends* still show varied

results, indicating their complexity. ChatGPT 4.5 excels in many categories, scoring 91.0 % in *ambiguity* and 95.5 % in *ldd & interrogatives*. *Punctuation* is the most difficult category for ChatGPT 4.5 achieving 76 % accuracy. One possible explanation is that GPT translations frequently include punctuation and other content not present in the original text (Hendy et al., 2023).

For **English–German**, the categories with the highest scores are *negation*, *verb tense/aspect/mood*, and *function word*. ChatGPT 4.5 performs well in *function word* (97.6 %) and *ldd & interrogatives*, although NLLBG still outperforms ChatGPT in *ldd & interrogatives*. ChatGPT and NMT models can improve in categories like *subordination* and *verb valency*, where scores are often below 90 %.

For **English–Russian**, the category with the highest average score (89.4 %) is *punctuation*. Categories like *verb semantics* and *lexical Morphology* pose significant challenges. The categories with the lowest accuracy are *ambiguity* with 51.8 %, followed by *coordination & ellipsis*. However, ChatGPT 4.5 achieves the lowest results in the category *false friends* with 61.5 % accuracy. ChatGPT performs best in *function word* (93.1 %) and *verb tense/aspect/mood* (85.9 %). The most challenging phenomenon for ChatGPT is *verb semantics* with a score of 47.1 %.

5.3 Phenomenon-level analysis

For **German–English**, the phenomenon macro-average for ChatGPT is 91.5 % with over 40 phenomena reaching a 100 % accuracy. There are no phenomena that reach 100 % accuracy across all models but some of the easier phenomena for most models include *phrasal verb*, *sluicing*, *polar question*, *ditransitive future I*, *passive voice* and other. The phenomena with the lowest accuracies are *idioms*, *modal negated - pluperfect*, and *resultative predicates*. In terms of *idioms*, ChatGPT performs better than most systems with 57.9 % accuracy.

Table 2 contains example outputs from two different phenomena for German–English. The first example comes from the phenomenon *extended adjective construction*, a frequent construction in German grammar, where the adjective is modified prepositional phrases or attributes. This structure tends to complicate the syntactic structure, making MT more challenging. The first translation is incorrect as it doesn't accurately convey the meaning of the original sentence. The second translation accurately conveys the meaning of the original sentence and uses correct English grammar. The third translation is also inaccurate due to the wrong word order and the incorrect use of an adjective instead of an adverb.

Extended Adjective Construction	
Auf der anderen Straßenseite stand ein laut weinendes Kind.	
On the other side of the street was a noisy child.	fail
A child was crying loudly across the street.	pass
Across the street stood a loud crying child.	fail
Resultative Predicate	
Es regnete die Stühle nass.	
It rained wet the chairs.	fail
It rained and the chairs got wet.	pass
It had a wet effect on the chairs.	fail

Table 2: Examples of German–English linguistic phenomena with passing and failing MT outputs.

The second example contains a *resultative predicate*. The first translation is incorrect because it does not follow the correct word order in English. The word-to-word translation of the German sentence is taken too directly, resulting in an awkward and non-sensical English sentence. The second translation is correct. It accurately conveys the meaning of the original German sentence and uses a natural English construction to do so. The third translation is also incorrect as "having a wet effect" is not typically used to describe things that are "wet" or that "get wet".

For **English–German**, the phenomenon-level macro average is similarly high as for the other language direction with 93 %. The phenomena for which all systems reach near 100 % accuracy include *inversion*, *multiple connectors*, *pied-piping*, *prepositional mwe*, *substitution*, *adverbial clause* and others. Most of the phenomena achieve high accuracies over 85 %, with some exceptions including *stripping*, *topicalization*, *verb semantics*, *mediopassive voice*, and *noun formation(er)*.

Table 3 contains translation examples from English–German. The first example contains a *functional shift*. Functional shift, or conversion, is when a word switches from one word class, or part of speech without changing its form Cannon (1985). In the first output, we can observe a correct structural change with the use of a common German prepositional phrase. In the second output, however, the word "wassappieren" is not a valid German word, resulting in an incomprehensible translation. Similarly, the third translation is also

Functional Shift	
You can whatsapp me on this number.	
Sie können mich per Whatsapp unter dieser Nummer erreichen.	pass
Sie können mich auf dieser Nummer wassappieren.	fail
Du kannst mich auf dieser Nummer aufpassen.	fail
Semantic Roles	
The bike accident broke Sarah's arm.	
Der Fahrradunfall brach Sarah den Arm.	fail
Bei dem Fahrradunfall brach sich Sarah den Arm.	pass

Table 3: Examples of English–German linguistic phenomena with passing and failing MT outputs.

Compound	
He denies that racism is just skin-deep.	
Он отрицает, что расизм — это просто глубинка.	fail
Он отрицает, что расизм поверхностен.	pass
Он отрицает, что расизм сводится только к цвету кожи.	fail
Idiom	
When things look black, there's always a silver lining.	
Когда все выглядит мрачно, всегда есть луч надежды.	pass
Когда все выглядит черным, всегда есть серебряная подкладка.	fail
Когда все выглядит черным, всегда есть худ без добра.	fail

Table 4: Examples of English–Russian linguistic phenomena with passing and failing MT outputs.

not a valid German sentence, it introduces a different verb, "aufpassen", which means "to look after" and doesn't fit the original meaning of the sentence. The second example deals with the problem of *semantic roles* also known as *thematic relations*. English has a broad range of semantic roles in the subject position and while German also allows for non-agentive semantic roles to be expressed as subjects, it may be more restrictive than English. In the incorrect translation, the accident itself is depicted as the direct agent of the action, which is unusual for German. According to the accurate translation, which follows the typical German sentence form, "Sarah's arm broke as a result of the accident".

For **English–Russian**, the phenomenon level macro-average accuracy lies at 77 %. In this year's submission, the following phenomena reached 97-100 % accuracy: *prepositional mwe*, *contact clause*, *object clause*. The two phenomena reaching the lowest accuracies were *idioms* and *semantic roles* with less than 40 % averages. The low accuracy for *idioms* and *semantic roles* are not surprising as t expressions still cause translation errors across all language pairs. ChatGPT 4.5 performs as the fourth-best system in all the averages, showing the lowest result for *semantic roles* as well.

Table 4 covers translation examples in English–Russian. For instance, the translation of a problematic English *compound* "skin-deep" into Russian. The first translation "Он отрицает, что расизм — это просто глубинка" means in Russian "He denies that racism is just a small rural town." "Глубинка" does have the same root as the word "deep" in Russian but has a completely different meaning, which makes this translation incorrect. The second structure is correct as it uses the adjective "поверхностен" or "superficial". The third translation is also incorrect as it means "He denies that racism is only about skin color" and states that the issue of racism is related to skin color, which was not present in the test item. The second example comes from the phenomenon *idiom*. This example includes a very common English non-literal expression "silver lining" meaning that there might be a positive aspect to a situation that may initially appear depressing or hopeless. The first translation correctly interprets the English idiom using a popular expression in Russian, "луч надежды" (ray of hope), reflecting the idea that even in bad times, there is always hope for something positive. The second translation renders the idiom literally. The Russian phrase "серебряная подкладка"(silver underlay) is not commonly used and does not accurately express the original meaning. In the third translation, an appropriate Russian proverb "There is no bad without good" is used to convey the meaning, but there's an error in the Russian expression: instead of "худа", there is a non-existent word "худ", making this translation incorrect.

5.4 Comparison with previous years

The progress of the systems' accuracy for particular categories through the last years can be seen in Table 8 for German–English (since 2018), Table 9 for English–German (since 2021) and Table 10 for English–Russian (since 2022). The calculation has been done based on the common test items without warnings over the years. Compared to last year, the micro- and macro-average scores for the German-English systems included in the comparison have either shown very small improvement or remained the same. For English–German, 3 systems (Online-G, Y, and W) showed an im-

provement, which in some categories sums up to several percentage points. In English–Russian, 5 out of the 7 the systems (Online-A, G, W, Y, and PROMT) showed an improvement which averages to 1-5 %. Whereas we have little information about the development behind the online systems, we can assume that English–Russian is still in active development, English–German has undergone minor improvements, whereas there seems to have been no development for German-English.

Interestingly enough, the Lan-Bridge performance has gotten worse both in micro and macro averages compared to last year. The drop in performance is important in light of Lan-Bridge's own system description. Their approach in the WMT23 competition has been shaped by the shift towards large-scale models and lies on prompt-based experiments. To understand the specific reasons for Lan-Bridge's drop in performance, a detailed analysis of their models, data, experiment designs, and evaluation metrics would be necessary.

6 Conclusions and Outlook

This paper presents a fine-grained, linguistically motivated test suite to evaluate machine translation outputs. The test suite was applied to evaluate and compare the outputs of 37 machine translation systems in three different language pairs: German–English, English–German, and English–Russian.

While the evaluation showed high scores for all language pairs, there was a clear drop in accuracy when dealing with structurally different languages, such as English and Russian. For this language pair, ChatGPT's performance falls in the second significance cluster. Although we didn't observe a systematic significant difference between ChatGPT 4.5 and other systems, it is important to highlight that ChatGPT 4.5 shows competitive results in the context of our evaluation. This indicates that ChatGPT 4.5, a general model, remains competitive in MT and sometimes performs better than some specialized NMT systems. Nevertheless, many linguistic nuances still pose difficulties for these models, demonstrating the continuous need for study and improvement in the field of MT. In terms of linguistic coverage, the current test suite stands out as one of the most extensive available. The semi-automated approach offers a more effective, while still fine-grained analysis in comparison to a typical human evaluation. When paired with other automated metrics or MQM analysis, this method can be seen as a valuable addition offering deeper insights into translation quality. The test suite approach is also highly versatile, allowing for the analysis of various tasks performed by LLMs in different contexts.

Limitations

The current test suite, evolving since 2016, was originally designed to evaluate weaker MT systems and focused on simpler linguistic phenomena. While we've introduced complexity with multi-sentence test items and more intricate sentences, it could be done only for a handful of phenomena and sentences. There are other limitations to consider. Firstly, this analysis is mostly limited to a sentence-level analysis. Secondly, all phenomena and categories are treated equally, although they may vary in their complexity. As mentioned earlier, the current evaluation rules prioritize accuracy in translating specific linguistic phenomena, sometimes at the expense of overall natural fluency, resulting in technically correct but less fluent outputs. To address some of these limitations, we consider including a linguistic acceptability score and an inter-annotator agreement score in future evaluations.

Acknowledgements

This research was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft; DFG) through the project TextQ, by the German Federal Ministry of Education through the project SocialWear (grant num. 01IW20002). We would like to thank Hans Uszkoreit, Aljoscha Burchardt, Ursula Strohriegel, Renlong Ai and He Wang for their prior contributions to the creation of the test suite.

References

Eleftherios Avramidis and Vivien Macketanz. 2022. Linguistically motivated evaluation of machine translation metrics based on a challenge set. In *Proceedings of the Seventh Conference on Machine Translation*, pages 514–529, Abu Dhabi. Association for Computational Linguistics.

Eleftherios Avramidis, Vivien Macketanz, Arle Lommel, and Hans Uszkoreit. 2018. Fine-grained evaluation of quality estimation for machine translation based on a linguistically motivated test suite. In *Proceedings of the AMTA 2018 Workshop on Translation Quality*

Estimation and Automatic Post-Editing, pages 243–248, Boston, MA. Association for Machine Translation in the Americas.

Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, Aljoscha Burchardt, and Sebastian Möller. 2020. Fine-grained linguistic evaluation for state-of-the-art Machine Translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 346–356, Online. Association for Computational Linguistics.

Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, and Hans Uszkoreit. 2019. Linguistic Evaluation of German-English Machine Translation Using a Test Suite. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 445–454, Florence, Italy. Association for Computational Linguistics.

Ondřej Bojar, Jiří Mírovský, Kateřina Rysová, and Magdaléna Rysová. 2018. EvalD Reference-Less Discourse Evaluation for WMT18. In *Proceedings of the Third Conference on Machine Translation*, pages 545–549, Belgium, Brussels. Association for Computational Linguistics.

Aljoscha Burchardt, Vivien Macketanz, Jon Dehdari, Georg Heigold, Jan-Thorsten Peter, and Philip Williams. 2017. A Linguistic Evaluation of Rule-Based, Phrase-Based, and Neural MT Engines. *The Prague Bulletin of Mathematical Linguistics*, 108:159–170.

Franck Burlot, Yves Scherrer, Vinit Ravishankar, Ondřej Bojar, Stig-Arne Grönroos, Maarit Koponen, Tommi Nieminen, and François Yvon. 2018. The WMT'18 Morpheval test suites for English-Czech, English-German, English-Finnish and Turkish-English. In *Proceedings of the Third Conference on Machine Translation*, pages 550–564, Belgium, Brussels. Association for Computational Linguistics.

Garland Cannon. 1985. Functional shift in english.

Sheila Castilho, Clodagh Quinn Mallon, Rahel Meister, and Shengya Yue. 2023. Do online Machine Translation Systems Care for Context? What About a GPT Model? In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 393–417, Tampere, Finland.

Silvie Cinkova and Ondřej Bojar. 2018. Testsuite on Czech–English Grammatical Contrasts. In *Proceedings of the Third Conference on Machine Translation*, pages 565–575, Belgium, Brussels. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondrej Bojar. 2021. Results of the WMT21 Metrics Shared Task: Evaluating Metrics with Expert-based Human Evaluations on TED and News Domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online.

Liane Guillou and Christian Hardmeier. 2016. PROTEST: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 636–643, Portorož, Slovenia. European Language Resources Association (ELRA).

Liane Guillou, Christian Hardmeier, Ekaterina Lapshinova-Koltunski, and Sharid Loáiciga. 2018. A Pronoun Test Suite Evaluation of the English–German MT Systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation*, pages 576–583, Belgium, Brussels. Association for Computational Linguistics.

Ulrich Heid and Elke Hildenbrand. 1991. Some practical experience with the use of test suites for the evaluation of SYSTRAN. In *the Proceedings of the Evaluators' Forum, Les Rasses*. Citeseer.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine.

Margaret King and Kirsten Falkedal. 1990. Using test suites in evaluation of machine translation systems. In *Proceedings of the 13th conference on Computational Linguistics*, volume 2, pages 211–216, Morristown, NJ, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Christof Monz, Makoto Morishita, Murray Kenton, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23). In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Tomasz Limisiewicz, and Gabriel Stanovsky. 2020. Gender coreference and bias evaluation at wmt 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 357–364, Online. Association for Computational Linguistics.

Vivien Macketanz, Renlong Ai, Aljoscha Burchardt, and Hans Uszkoreit. 2018a. TQ-AutoTest – an automated test suite for (machine) translation quality. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, and Hans Uszkoreit. 2018b. Fine-grained evaluation of German-English machine translation based on a test suite. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 578–587, Belgium, Brussels. Association for Computational Linguistics.

Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, and Hans Uszkoreit. 2018c. Fine-grained evaluation of German-English Machine Translation based on a Test Suite. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 584–593, Belgium, Brussels. Association for Computational Linguistics.

Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, He Wang, Renlong Ai, Shushen Manakhimova, Ursula Strohriegel, Sebastian Möller, and Hans Uszkoreit. 2022a. A Linguistically Motivated Test Suite to Semi-Automatically Evaluate German–English Machine Translation Output. In *Proceedings of the Language Resources and Evaluation Conference*, pages 936–947, Marseille, France. European Language Resources Association.

Vivien Macketanz, Eleftherios Avramidis, Shushen Manakhimova, and Sebastian Möller. 2021. Linguistic Evaluation for the 2021 State-of-the-art Machine Translation Systems for German to English and English to German. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1059–1073, Online. Association for Computational Linguistics.

Vivien Macketanz, Shushen Manakhimova, Eleftherios Avramidis, Ekaterina Lapshinova-koltunski, Sergei Bagdasarov, and Sebastian Möller. 2022b. Linguistically motivated evaluation of the 2022 state-of-the-art machine translation systems for three language directions. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 432–449, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Maja Popović. 2019. Evaluating Conjunction Disambiguation on English-to-German and French-to-German WMT 2019 Translation Hypotheses. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 464–469, Florence, Italy. Association for Computational Linguistics.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2019. The MuCoW Test Suite at WMT 2019: Automatically Harvested Multilingual Contrastive Word Sense Disambiguation Test Sets for Machine Translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 470–480, Florence, Italy. Association for Computational Linguistics.

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan Awadalla. 2023. Do gpts produce less literal translations?

Annette Rios, Mathias Müller, and Rico Sennrich. 2018. The Word Sense Disambiguation Test Suite at WMT18. In *Proceedings of the Third Conference on Machine Translation*, pages 594–602, Belgium, Brussels. Association for Computational Linguistics.

Kateřina Rysová, Magdaléna Rysová, Tomáš Musil, Lucie Poláková, and Ondřej Bojar. 2019. A Test Suite and Manual Evaluation of Document-Level NMT at WMT19. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 455–463, Florence, Italy. Association for Computational Linguistics.

Yves Scherrer, Alessandro Raganato, and Jörg Tiedemann. 2020. The MUCOW word sense disambiguation test suite at WMT 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 365–370, Online. Association for Computational Linguistics.

Kirill Semenov and Ondřej Bojar. 2022. Automated evaluation metric for terminology consistency in MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 450–457, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tereza Vojtěchová, Michal Novák, Miloš Klouček, and Ondřej Bojar. 2019. SAO WMT19 Test Suite: Machine Translation of Audit Reports. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 481–493, Florence, Italy. Association for Computational Linguistics.

Andrew Way. 1991. Developer-Oriented Evaluation of MT Systems. In *Proceedings of the Evaluators' Forum*, pages 237–244, Les Rasses, Vaud, Switzerland. ISSCO.

Marion Weller-di Marco and Alexander Fraser. 2022. Test suite evaluation: Morphological challenges and pronoun translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 458–468, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yangjian Wu and Gang Hu. 2023. Exploring Prompt Engineering with GPT Language Models for Document-Level Machine Translation: Insights and Findings. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Vilém Zouhar, Tereza Vojtěchová, and Ondřej Bojar. 2020. WMT20 Document-Level Markable Error Exploration. In *Proceedings of the Fifth Conference on*

Machine Translation, pages 371–380, Online. Association for Computational Linguistics.

A Analysis based on categories

categ	count	Onl-W	Onl-A	Onl-B	ChatG	Onl-M	Onl-Y	NLLBM	NLLBG	Onl-G	LanBr	GTCOM	ZengH	AIRC	avg
Ambiguity	78	**85.9**	**88.5**	**93.6**	**91.0**	84.6	**87.2**	**87.2**	84.6	**87.2**	78.2	75.6	**88.5**	62.8	84.2
Composition	45	**100.0**	**100.0**	**97.8**	**100.0**	**97.8**	**100.0**	93.3	**95.6**	**95.6**	91.1	**95.6**	**95.6**	77.8	95.4
Coordination & ellipsis	49	**93.9**	**93.9**	**91.8**	**89.8**	77.6	**91.8**	**85.7**	**83.7**	**93.9**	77.6	**91.8**	**87.8**	81.6	87.8
False friends	36	**91.7**	**86.1**	**77.8**	**83.3**	**83.3**	69.4	**83.3**	**80.6**	**80.6**	75.0	75.0	72.2	52.8	77.8
Function word	61	**90.2**	**93.4**	**93.4**	**91.8**	**91.8**	**88.5**	**95.1**	**91.8**	**90.2**	78.7	83.6	52.5	65.6	85.1
LDD & interrogatives	154	87.0	90.3	88.3	**95.5**	87.7	87.7	87.0	89.6	90.3	79.2	85.1	72.1	66.2	85.1
MWE	76	**90.8**	**82.9**	**82.9**	**88.2**	77.6	80.3	**81.6**	**82.9**	80.3	71.1	76.3	**84.2**	53.9	79.5
Named entity & terminology	20	**95.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**95.0**	50.0	0.0	**90.0**	86.9
Negation	19	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	94.7	99.6
Non-verbal agreement	60	**93.3**	90.0	**96.7**	**93.3**	**95.0**	**98.3**	**96.7**	**96.7**	88.3	86.7	81.7	**95.0**	71.7	91.0
Punctuation	50	**100.0**	**100.0**	94.0	76.0	**100.0**	74.0	74.0	74.0	64.0	84.0	70.0	50.0	94.0	81.1
Subordination	158	**91.1**	89.2	**92.4**	**91.8**	**92.4**	**93.7**	**94.9**	**93.0**	**92.4**	75.9	86.7	85.4	76.6	88.9
Verb tense/aspect/mood	2347	**93.7**	**94.0**	91.4	**92.9**	88.0	**94.0**	84.3	84.4	**93.1**	81.8	**93.8**	**93.8**	86.6	90.1
Verb valency	81	**84.0**	**84.0**	**85.2**	**88.9**	**85.2**	**84.0**	**85.2**	**87.7**	77.8	77.8	**82.7**	**81.5**	65.4	82.2
micro-average	3234	**92.9**	**93.0**	91.2	**92.5**	88.3	**92.5**	85.6	85.6	91.5	81.2	90.7	89.4	82.2	89.0
macro-average	3234	**92.6**	**92.3**	**91.8**	**91.6**	90.1	89.2	89.2	88.9	88.1	82.3	82.0	75.6	74.3	86.8

Table 5: Accuracies (%) of successful translations on the category level for German–English. Boldface indicates the significantly best performing systems per row.

categ	count	ChatG	Onl-W	Onl-B	Onl-A	Onl-Y	NLLBG	Onl-G	NLLBM	Onl-M	ZengH	LanBr	AIRC	avg
Ambiguity	24	**95.8**	**95.8**	**91.7**	**87.5**	**83.3**	**83.3**	**87.5**	**83.3**	**87.5**	**91.7**	75.0	50.0	84.4
Coordination & ellipsis	74	**90.5**	78.4	**93.2**	**85.1**	**93.2**	70.3	**90.5**	67.6	82.4	74.3	71.6	63.5	80.1
False friends	33	**93.9**	**93.9**	**97.0**	**93.9**	**93.9**	**97.0**	**90.9**	**97.0**	**93.9**	**93.9**	**93.9**	81.8	93.4
Function word	41	**97.6**	**97.6**	**97.6**	**97.6**	**97.6**	**97.6**	**97.6**	**97.6**	**97.6**	75.6	**97.6**	85.4	94.7
LDD & interrogatives	131	**96.9**	**96.2**	**95.4**	**96.9**	**96.9**	**94.7**	**93.9**	**93.9**	**93.9**	88.5	**92.4**	84.7	93.7
Lexical Morphology	28	**85.7**	**85.7**	**82.1**	**75.0**	**67.9**	**67.9**	**64.3**	**64.3**	57.1	**82.1**	42.9	25.0	66.7
MWE	95	**95.8**	**97.9**	**96.8**	91.6	**95.8**	85.3	89.5	86.3	86.3	92.6	78.9	68.4	88.8
Named entity & terminology	73	**95.9**	**95.9**	**95.9**	**97.3**	**97.3**	83.6	**94.5**	87.7	**94.5**	87.7	78.1	90.4	91.6
Negation	13	100.0	100.0	100.0	100.0	100.0	92.3	100.0	92.3	100.0	100.0	100.0	100.0	98.7
Non-verbal agreement	90	**97.8**	**94.4**	90.0	88.9	92.2	**94.4**	**93.3**	**95.6**	**95.6**	92.2	87.8	74.4	91.4
Punctuation	36	83.3	**97.2**	80.6	**88.9**	77.8	80.6	80.6	**86.1**	83.3	61.1	80.6	72.2	81.0
Subordination	136	**99.3**	**97.1**	**97.8**	**97.8**	**96.3**	**97.8**	**97.8**	**97.8**	**97.8**	**97.1**	**99.3**	92.6	97.4
Verb semantics	4	75.0	75.0	75.0	50.0	50.0	100.0	50.0	75.0	50.0	50.0	50.0	25.0	60.4
Verb tense/aspect/mood	2237	99.1	98.4	98.7	99.0	**99.6**	97.0	99.1	97.1	98.4	99.2	97.2	91.6	97.9
Verb valency	94	**86.2**	**86.2**	**88.3**	**86.2**	**79.8**	77.7	76.6	**80.9**	**78.7**	**86.2**	72.3	59.6	79.9
micro-average	3109	**97.8**	97.0	**97.2**	97.0	**97.4**	94.4	96.6	94.7	95.9	95.9	93.5	87.1	95.4

categ	count	ChatG	Onl-W	Onl-B	Onl-A	Onl-Y	NLLBG	Onl-G	NLLBM	Onl-M	ZengH	LanBr	AIRC	avg
macro-average	3109	**92.9**	**92.6**	**92.0**	89.0	88.1	88.0	87.1	86.8	86.5	84.8	81.2	71.0	86.7

Table 6: Accuracies (%) of successful translations on the category level for English–German. Boldface indicates the significantly best performing systems per row.

categ	count	Onl-G	Onl-W	Onl-B	ChatG	Onl-Y	Onl-A	NLLBM	NLLBG	Onl-M	PROMT	ZengH	LanBr	avg
Ambiguity	20	**70.0**	**60.0**	50.0	**85.0**	55.0	45.0	50.0	50.0	35.0	30.0	45.0	25.0	50.0
Coordination & ellipsis	89	**82.0**	**83.1**	67.4	**77.5**	68.5	65.2	62.9	66.3	67.4	58.4	50.6	49.4	66.6
False friends	14	85.7	85.7	78.6	64.3	85.7	71.4	57.1	64.3	64.3	57.1	71.4	50.0	69.6
Function word	29	**96.6**	**96.6**	**96.6**	**93.1**	**82.8**	**82.8**	**96.6**	**93.1**	**96.6**	**86.2**	37.9	75.9	86.2
LDD & interrogatives	61	**95.1**	**95.1**	**91.8**	**88.5**	**93.4**	**91.8**	**88.5**	**88.5**	85.2	85.2	73.8	78.7	88.0
Lexical Morphology	29	**86.2**	**86.2**	**75.9**	**86.2**	**65.5**	62.1	62.1	**65.5**	41.4	51.7	58.6	55.2	66.4
MWE	71	**76.1**	**73.2**	**76.1**	**70.4**	59.2	**69.0**	**67.6**	**66.2**	**60.6**	**60.6**	**69.0**	54.9	66.9
Named entity & terminology	71	**87.3**	**77.5**	**81.7**	73.2	69.0	**76.1**	63.4	63.4	69.0	59.2	**80.3**	60.6	71.7
Negation	4	75.0	100.0	100.0	75.0	100.0	75.0	75.0	75.0	100.0	75.0	100.0	50.0	83.3
Non-verbal agreement	80	**76.3**	**86.3**	**75.0**	**82.5**	73.8	72.5	**81.3**	**81.3**	73.8	**75.0**	66.3	65.0	75.7
Punctuation	12	**100.0**	**83.3**	**91.7**	66.7	75.0	**100.0**	**83.3**	**83.3**	66.7	**91.7**	0.0	**91.7**	77.8
Subordination	130	**93.8**	**96.9**	**93.8**	**93.8**	**93.8**	90.0	86.9	88.5	**93.8**	89.2	68.5	83.1	89.4
Verb semantics	17	**94.1**	**82.4**	**76.5**	47.1	58.8	**76.5**	52.9	47.1	58.8	58.8	58.8	41.2	62.7
Verb tense/aspect/mood	156	**91.7**	**94.2**	85.9	85.9	87.2	87.8	84.0	82.1	83.3	84.0	66.7	75.0	84.0
Verb valency	126	**84.9**	**81.7**	**79.4**	**78.6**	**77.0**	72.2	68.3	64.3	73.0	70.6	69.8	60.3	73.3
micro-average	909	**86.9**	**86.8**	81.7	81.7	78.3	78.0	75.2	74.8	75.4	72.9	65.0	65.7	76.9
macro-average	909	**86.3**	**85.5**	81.3	77.9	76.3	75.8	72.0	71.9	71.3	68.9	61.1	61.1	74.1

Table 7: Accuracies (%) of successful translations on the category level for English–Russian. Boldface indicates the significantly best-performing systems per row.

B Comparison through the years

		Lan-Bridge		online-A						online-B						online-G						online-W			online-Y				
category	count	22	23	18	19	20	21	22	23	18	19	20	21	22	23	18	19	20	21	22	23	21	22	23	18	19	21	22	23
Ambiguity	75	89	80	69	71	77	83	84	88	76	77	79	85	93	93	72	75	84	85	88	88	85	84	87	68	79	83	84	88
Composition	45	98	91	82	93	96	96	96	100	98	98	98	100	98	98	73	87	98	98	98	96	96	96	100	91	93	93	100	100
Coordination & ellipsis	28	86	71	86	86	86	89	89	89	86	86	89	89	93	89	50	64	75	89	89	89	89	89	89	86	86	89	89	89
False friends	36	83	75	72	72	69	83	83	86	75	78	81	75	78	78	72	72	78	81	83	81	86	81	92	67	92	75	78	69
Function word	58	95	81	84	90	88	91	91	95	78	78	93	88	95	95	50	93	93	95	93	91	95	93	90	91	91	84	88	91
LDD & interrogatives	72	92	81	79	75	85	89	89	93	85	85	89	94	92	92	64	72	92	88	89	90	89	90	92	83	79	90	85	89
MWE	64	77	73	67	67	73	81	83	83	73	73	78	78	80	80	67	69	81	81	81	81	89	89	91	72	73	75	83	81
Named entity & terminology	9	100	100	100	100	100	100	100	100	100	100	78	100	100	100	100	100	100	100	100	100	100	100	89	89	100	100	100	100
Negation	16	100	100	100	100	100	100	100	100	94	94	100	100	100	100	63	100	100	100	100	100	100	100	100	100	100	100	100	100

		Lan-Bridge		online-A						online-B						online-G						online-W			online-Y				
category	count	22	23	18	19	20	21	22	23	18	19	20	21	22	23	18	19	20	21	22	23	21	22	23	18	19	21	22	23
Non-verbal agreement	55	98	85	78	84	84	93	93	91	87	87	87	98	96	96	58	82	91	91	89	89	96	96	95	80	82	85	93	96
Punctuation	33	94	94	100	100	100	100	100	100	97	97	97	100	94	94	88	88	91	91	91	91	97	100	100	100	100	100	100	97
Subordination	87	94	76	94	79	94	94	94	95	87	89	94	95	97	97	82	90	93	91	94	93	93	92	95	93	93	93	94	95
Verb tense/aspect/mood	2775	87	79	80	89	86	90	90	90	82	82	84	83	86	86	55	74	88	84	89	90	90	88	92	78	80	81	88	89
Verb valency	56	88	84	79	84	88	88	88	88	82	82	91	89	88	88	71	79	88	88	84	84	89	89	88	80	82	82	86	86
micro-average	3409	88	79	80	88	86	90	90	90	83	83	85	85	87	87	57	76	88	85	89	90	90	88	92	79	81	82	89	89
macro-average	3409	91	84	84	85	88	91	91	93	86	86	88	91	92	92	69	82	89	90	91	90	93	92	93	84	88	88	91	91

Table 8: Comparisons of the accuracy (%) of several German–English systems through the years.

		Lan-Bridge		online-A			online-B			online-G			online-W			online-Y		
category	count	22	23	21	22	23	21	22	23	21	22	23	21	22	23	21	22	23
Ambiguity	24	83.3	75.0	91.7	87.5	87.5	91.7	91.7	91.7	75.0	83.3	87.5	95.8	95.8	95.8	70.8	79.2	83.3
Coordination & ellipsis	68	86.8	63.2	70.6	82.4	79.4	82.4	88.2	88.2	73.5	85.3	88.2	66.2	67.6	69.1	67.6	76.5	86.8
False friends	36	86.1	86.1	86.1	86.1	88.9	83.3	88.9	91.7	83.3	91.7	83.3	88.9	91.7	91.7	86.1	86.1	86.1
Function word	39	97.4	97.4	97.4	97.4	97.4	100.0	97.4	97.4	97.4	97.4	97.4	100.0	100.0	97.4	97.4	97.4	97.4
MWE	96	87.5	81.3	86.5	89.6	91.7	92.7	95.8	96.9	81.3	89.6	90.6	93.8	97.9	97.9	80.2	85.4	95.8
Named entity & terminology	64	98.4	79.7	96.9	96.9	96.9	93.8	100.0	96.9	81.3	93.8	95.3	98.4	95.3	96.9	93.8	96.9	98.4
Negation	15	100.0	93.3	93.3	100.0	100.0	93.3	100.0	93.3	93.3	93.3	93.3	100.0	100.0	100.0	100.0	100.0	100.0
Non-verbal agreement	64	98.4	96.9	96.9	96.9	96.9	96.9	96.9	96.9	95.3	98.4	98.4	95.3	96.9	95.3	95.3	95.3	98.4
Punctuation	18	66.7	66.7	94.4	94.4	83.3	66.7	66.7	66.7	50.0	66.7	66.7	94.4	88.9	94.4	66.7	66.7	66.7
Subordination	129	98.4	99.2	98.4	98.4	97.7	97.7	98.4	97.7	93.8	99.2	97.7	97.7	97.7	98.4	94.6	93.8	96.1
Verb tense/aspect/mood	2526	99.3	97.3	96.1	98.6	98.7	99.1	98.8	98.4	94.9	97.6	98.9	96.8	96.6	98.1	93.0	96.6	99.6
Verb valency	76	88.2	78.9	84.2	86.8	93.4	88.2	89.5	92.1	77.6	88.2	85.5	89.5	88.2	90.8	80.3	85.5	86.8
micro-average	3155	97.9	94.9	94.9	97.4	97.5	97.7	97.9	97.6	92.8	96.5	97.4	95.9	95.8	97.1	91.6	95.0	98.2
macro-average	3155	90.9	84.6	91.0	92.9	92.6	90.5	92.7	92.3	83.1	90.4	90.2	93.1	93.1	93.8	85.5	88.3	91.3

Table 9: Comparisons of the accuracy (%) of several English–German systems through the years.

		Lan-Bridge		online-A		online-B		online-G		online-W		online-Y		PROMT	
category	count	22	23	22	23	22	23	22	23	22	23	22	23	22	23
Ambiguity	7	71.0	57.0	71.0	86.0	86.0	86.0	86.0	86.0	100.0	86.0	71.0	86.0	57.0	71.0
Coordination & ellipsis	30	50.0	40.0	43.0	60.0	57.0	57.0	80.0	80.0	73.0	77.0	53.0	60.0	60.0	53.0
False friends	5	60.0	80.0	60.0	60.0	80.0	80.0	80.0	80.0	80.0	80.0	60.0	80.0	60.0	60.0
Function word	10	80.0	60.0	80.0	90.0	90.0	90.0	90.0	90.0	90.0	100.0	90.0	70.0	80.0	80.0
MWE	32	63.0	59.0	63.0	66.0	75.0	75.0	69.0	72.0	75.0	75.0	66.0	66.0	63.0	66.0
Named entity & terminology	22	82.0	64.0	68.0	77.0	86.0	86.0	91.0	95.0	77.0	68.0	73.0	77.0	73.0	68.0

category	count	Lan-Bridge 22	Lan-Bridge 23	online-A 22	online-A 23	online-B 22	online-B 23	online-G 22	online-G 23	online-W 22	online-W 23	online-Y 22	online-Y 23	PROMT 22	PROMT 23
Negation	4	100.0	50.0	75.0	75.0	100.0	100.0	75.0	75.0	100.0	100.0	100.0	100.0	75.0	75.0
Non-verbal agreement	10	80.0	50.0	80.0	80.0	90.0	90.0	80.0	80.0	80.0	90.0	70.0	80.0	80.0	80.0
Punctuation	5	80.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	80.0	80.0	100.0	80.0
Subordination	48	90.0	81.0	81.0	83.0	92.0	92.0	90.0	92.0	92.0	98.0	79.0	96.0	81.0	85.0
Verb tense/aspect/mood	61	77.0	75.0	77.0	82.0	75.0	79.0	77.0	84.0	75.0	89.0	70.0	74.0	77.0	82.0
Verb valency	30	73.0	53.0	83.0	80.0	77.0	77.0	87.0	83.0	90.0	83.0	77.0	77.0	73.0	80.0
micro-average	264	75.0	65.0	72.0	77.0	80.0	80.0	82.0	84.0	82.0	86.0	72.0	77.0	73.0	75.0
macro-average	264	75.0	64.0	74.0	78.0	84.0	84.0	84.0	85.0	86.0	87.0	74.0	79.0	73.0	73.0

Table 10: Comparisons of the accuracy (%) of several German–English systems through the years.

C Detailed analysis on a phenomenon-level

categ	count	Onl-W	Onl-A	Onl-B	ChatG	Onl-M	Onl-Y	NLLBM	NLLBG	Onl-G	LanBr	GTCOM	ZengH	AIRC	avg
Ambiguity	78	**85.9**	**88.5**	**93.6**	**91.0**	84.6	**87.2**	**87.2**	84.6	**87.2**	78.2	75.6	**88.5**	62.8	84.2
Lexical ambiguity	62	**91.9**	**93.5**	**95.2**	**90.3**	85.5	**87.1**	**90.3**	85.5	**88.7**	80.6	79.0	**90.3**	67.7	86.6
Structural ambiguity	16	62.5	**68.8**	**87.5**	**93.8**	**81.3**	**87.5**	**75.0**	**81.3**	**81.3**	**68.8**	62.5	**81.3**	43.8	75.0
Composition	45	**100.0**	**100.0**	**97.8**	**100.0**	**97.8**	**100.0**	93.3	**95.6**	**95.6**	91.1	**95.6**	**95.6**	77.8	95.4
Compound	26	**100.0**	**100.0**	**100.0**	**100.0**	**96.2**	**100.0**	88.5	**92.3**	**96.2**	84.6	**92.3**	**96.2**	76.9	94.1
Phrasal verb	19	**100.0**	**100.0**	**94.7**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**94.7**	**100.0**	**100.0**	**94.7**	78.9	97.2
Coordination & ellipsis	49	**93.9**	**93.9**	**91.8**	**89.8**	77.6	**91.8**	**85.7**	**83.7**	**93.9**	77.6	**91.8**	**87.8**	81.6	87.8
Gapping	19	**100.0**	**100.0**	**100.0**	**89.5**	**94.7**	**94.7**	**89.5**	**89.5**	**100.0**	73.7	**94.7**	**94.7**	**89.5**	93.1
Right node raising	18	**83.3**	**83.3**	**83.3**	**83.3**	50.0	**83.3**	**72.2**	**66.7**	**83.3**	**66.7**	**83.3**	**88.9**	**61.1**	76.1
Sluicing	12	**100.0**	**100.0**	**91.7**	**100.0**	**91.7**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	75.0	**100.0**	96.8
False friends	36	**91.7**	**86.1**	**77.8**	**83.3**	**83.3**	69.4	**83.3**	**80.6**	**80.6**	75.0	75.0	72.2	52.8	77.8
Function word	61	**90.2**	**93.4**	**93.4**	**91.8**	**91.8**	**88.5**	**95.1**	**91.8**	**90.2**	78.7	83.6	52.5	65.6	85.1
Focus particle	22	**95.5**	**100.0**	**100.0**	**100.0**	**100.0**	**95.5**	**95.5**	**90.9**	**100.0**	**90.9**	**90.9**	**95.5**	81.8	95.1
Modal particle	20	**80.0**	**85.0**	**80.0**	**75.0**	**75.0**	**75.0**	**90.0**	**85.0**	**70.0**	50.0	**70.0**	40.0	**70.0**	72.7
Question tag	19	**94.7**	**94.7**	**100.0**	**100.0**	**100.0**	**94.7**	**100.0**	**100.0**	**100.0**	**94.7**	**89.5**	15.8	42.1	86.6
LDD & interrogatives	154	87.0	90.3	88.3	**95.5**	87.7	87.7	87.0	89.6	90.3	79.2	85.1	72.1	66.2	85.1
Extended adjective construction	14	**100.0**	**92.9**	**100.0**	**85.7**	**92.9**	**92.9**	78.6	78.6	**100.0**	**92.9**	**92.9**	**92.9**	78.6	90.7
Extraposition	18	72.2	83.3	61.1	83.3	77.8	77.8	83.3	88.9	66.7	72.2	72.2	72.2	66.7	75.2
Multiple connectors	19	84.2	78.9	**89.5**	**100.0**	73.7	78.9	78.9	78.9	84.2	**89.5**	**89.5**	**89.5**	78.9	84.2
Pied-piping	20	85.0	**90.0**	**90.0**	**100.0**	**95.0**	**90.0**	**90.0**	**95.0**	**90.0**	75.0	80.0	80.0	60.0	86.2
Polar question	20	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	70.0	**100.0**	25.0	75.0	90.0
Scrambling	15	**86.7**	**93.3**	**93.3**	**93.3**	**93.3**	**86.7**	**93.3**	**93.3**	**93.3**	**66.7**	60.0	**86.7**	33.3	82.6
Topicalization	17	58.8	**76.5**	**70.6**	**94.1**	**76.5**	**76.5**	**76.5**	**82.4**	**88.2**	**70.6**	**82.4**	64.7	41.2	73.8
Wh-movement	31	**100.0**	**100.0**	**96.8**	**100.0**	90.3	**93.5**	90.3	**93.5**	**96.8**	90.3	**93.5**	74.2	80.6	92.3
MWE	76	**90.8**	**82.9**	**82.9**	**88.2**	77.6	80.3	**81.6**	**82.9**	80.3	71.1	76.3	**84.2**	53.9	79.5

categ	count	Onl-W	Onl-A	Onl-B	ChatG	Onl-M	Onl-Y	NLLBM	NLLBG	Onl-G	LanBr	GTCOM	ZengH	AIRC	avg
Collocation	19	**100.0**	**100.0**	**100.0**	**100.0**	**94.7**	**100.0**	**100.0**	**94.7**	**100.0**	84.2	**89.5**	**100.0**	57.9	93.9
Idiom	19	**63.2**	**31.6**	**42.1**	**57.9**	**15.8**	**21.1**	**31.6**	**36.8**	**26.3**	**10.5**	**15.8**	**36.8**	0.0	30.0
Prepositional MWE	19	**100.0**	**100.0**	**89.5**	**94.7**	**100.0**	**100.0**	**94.7**	**100.0**	**100.0**	**94.7**	**100.0**	**100.0**	68.4	95.5
Verbal MWE	19	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	94.7	94.7	100.0	100.0	89.5	98.4
Named entity & terminology	20	**95.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**95.0**	50.0	0.0	**90.0**	86.9
Date	20	**95.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**95.0**	50.0	0.0	**90.0**	86.9
Negation	19	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	94.7	99.6
Non-verbal agreement	60	**93.3**	90.0	**96.7**	**93.3**	**95.0**	**98.3**	**96.7**	**96.7**	88.3	86.7	81.7	**95.0**	71.7	91.0
Coreference	19	**94.7**	84.2	**89.5**	**94.7**	**89.5**	**94.7**	**94.7**	**94.7**	78.9	78.9	68.4	**100.0**	63.2	86.6
External possessor	21	**90.5**	**90.5**	**100.0**	**90.5**	**95.2**	**100.0**	**95.2**	**95.2**	**90.5**	**95.2**	81.0	**90.5**	57.1	90.1
Internal possessor	20	**95.0**	**95.0**	**100.0**	**95.0**	**100.0**	**100.0**	**100.0**	**100.0**	**95.0**	85.0	**95.0**	**95.0**	**95.0**	96.2
Punctuation	50	**100.0**	**100.0**	94.0	76.0	**100.0**	74.0	74.0	74.0	64.0	84.0	70.0	50.0	94.0	81.1
Comma	19	100.0	100.0	100.0	94.7	100.0	94.7	100.0	100.0	100.0	94.7	94.7	100.0	94.7	98.0
Quotation marks	31	**100.0**	**100.0**	90.3	64.5	**100.0**	61.3	58.1	58.1	41.9	77.4	54.8	19.4	**93.5**	70.7
Subordination	158	**91.1**	89.2	**92.4**	**91.8**	**92.4**	**93.7**	**94.9**	**93.0**	**92.4**	75.9	86.7	85.4	76.6	88.9
Adverbial clause	20	**90.0**	**90.0**	**100.0**	**90.0**	**90.0**	**95.0**	**95.0**	**90.0**	**90.0**	75.0	85.0	**90.0**	**90.0**	90.0
Cleft sentence	20	**95.0**	**95.0**	**95.0**	**95.0**	**95.0**	**100.0**	**95.0**	**95.0**	**100.0**	60.0	**90.0**	**95.0**	70.0	90.8
Free relative clause	14	100.0	92.9	92.9	100.0	85.7	100.0	100.0	85.7	100.0	100.0	100.0	85.7	92.9	95.1
Indirect speech	15	**86.7**	80.0	**93.3**	**86.7**	**100.0**	**93.3**	**100.0**	**100.0**	**93.3**	60.0	66.7	80.0	66.7	85.1
Infinitive clause	19	100.0	94.7	94.7	100.0	100.0	94.7	100.0	100.0	100.0	89.5	100.0	94.7	89.5	96.8
Object clause	16	**100.0**	**100.0**	**93.8**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**87.5**	**93.8**	**93.8**	81.3	96.2
Pseudo-cleft sentence	18	**77.8**	**83.3**	**83.3**	**83.3**	**72.2**	**83.3**	**72.2**	**72.2**	**72.2**	**66.7**	**66.7**	**61.1**	27.8	70.9
Relative clause	18	83.3	77.8	83.3	83.3	**88.9**	77.8	**100.0**	**100.0**	83.3	83.3	83.3	83.3	83.3	85.5
Subject clause	18	**88.9**	**88.9**	**94.4**	**88.9**	**100.0**	**100.0**	**94.4**	**94.4**	**94.4**	66.7	**94.4**	83.3	**88.9**	90.6
Verb tense/aspect/mood	2347	**93.7**	**94.0**	91.4	**92.9**	88.0	**94.0**	84.3	84.4	**93.1**	81.8	**93.8**	**93.8**	86.6	90.1
Conditional	16	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	93.8	99.5
Ditransitive - future I	36	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	91.7	**100.0**	**100.0**	**100.0**	99.4
Ditransitive - future I subjunctive II	24	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	87.5	**100.0**	**100.0**	**100.0**	99.0
Ditransitive - future II	31	**100.0**	**96.8**	**100.0**	83.9	51.6	**100.0**	32.3	25.8	**100.0**	80.6	**100.0**	**100.0**	67.7	79.9
Ditransitive - future II subjunctive II	31	**93.5**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**96.8**	**100.0**	**100.0**	87.1	98.3
Ditransitive - perfect	35	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**97.1**	**97.1**	**100.0**	88.6	**100.0**	**100.0**	**97.1**	98.5
Ditransitive - pluperfect	29	**100.0**	89.7	58.6	75.9	10.3	**93.1**	31.0	34.5	65.5	75.9	**93.1**	**96.6**	75.9	69.2
Ditransitive - pluperfect subjunctive II	25	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Ditransitive - present	24	**91.7**	**95.8**	**100.0**	**100.0**	**100.0**	**95.8**	**91.7**	87.5	87.5	83.3	**100.0**	**100.0**	**95.8**	94.6
Ditransitive - preterite	31	**100.0**	**93.5**	**93.5**	**96.8**	90.3	90.3	**96.8**	**96.8**	83.9	74.2	87.1	**96.8**	77.4	90.6
Ditransitive - preterite subjunctive II	26	**92.3**	88.5	80.8	88.5	**96.2**	84.6	**100.0**	**100.0**	84.6	80.8	**96.2**	80.8	76.9	88.5
Imperative	19	**100.0**	**100.0**	**100.0**	**100.0**	**89.5**	**94.7**	**100.0**	**94.7**	**94.7**	63.2	**89.5**	**89.5**	78.9	91.9
Intransitive - future I	32	**96.9**	**96.9**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	68.8	**96.9**	87.5	**96.9**	95.7
Intransitive - future I subjunctive II	29	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	89.7	**100.0**	**100.0**	**100.0**	99.2
Intransitive - future II	31	**100.0**	90.3	**96.8**	74.2	61.3	**100.0**	51.6	54.8	**96.8**	58.1	29.0	**100.0**	90.3	77.2
Intransitive - future II subjunctive II	7	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	14.3	100.0	100.0	93.4
Intransitive - perfect	76	**100.0**	**100.0**	**100.0**	**100.0**	**97.4**	**100.0**	94.7	92.1	**100.0**	60.5	**100.0**	**98.7**	92.1	95.0
Intransitive - pluperfect	32	**90.6**	**90.6**	**84.4**	**96.9**	28.1	**96.9**	25.0	25.0	68.8	37.5	**93.8**	**96.9**	**84.4**	70.7

categ	count	Onl-W	Onl-A	Onl-B	ChatG	Onl-M	Onl-Y	NLLBM	NLLBG	Onl-G	LanBr	GTCOM	ZengH	AIRC	avg
Intransitive - pluperfect subjunctive II	15	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	60.0	**100.0**	**100.0**	80.0	95.4
Intransitive - present	31	90.3	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	54.8	**100.0**	**100.0**	**100.0**	95.8
Intransitive - preterite	55	92.7	94.5	94.5	**100.0**	92.7	**100.0**	**96.4**	94.5	94.5	52.7	**96.4**	85.5	83.6	90.6
Intransitive - preterite subjunctive II	19	**57.9**	**63.2**	**78.9**	**84.2**	**63.2**	**78.9**	**68.4**	**68.4**	**68.4**	21.1	**73.7**	**57.9**	**57.9**	64.8
Modal - future I	95	**100.0**	**100.0**	**100.0**	**98.9**	**100.0**	89.5	96.8	93.7	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	98.4
Modal - future I subjunctive II	59	91.5	94.9	88.1	64.4	64.4	88.1	57.6	59.3	93.2	91.5	**100.0**	89.8	83.1	82.0
Modal - perfect	78	**78.2**	**78.2**	**76.9**	69.2	**74.4**	**82.1**	**84.6**	**78.2**	**79.5**	55.1	**83.3**	**83.3**	41.0	74.2
Modal - pluperfect	37	**86.5**	45.9	16.2	32.4	10.8	67.6	8.1	10.8	56.8	40.5	**75.7**	51.4	54.1	42.8
Modal - pluperfect subjunctive II	46	**73.9**	**71.7**	**73.9**	**76.1**	**69.6**	**71.7**	45.7	52.2	**69.6**	**73.9**	**76.1**	**73.9**	54.3	67.9
Modal - present	109	93.6	94.5	92.7	**100.0**	85.3	89.9	78.9	84.4	89.9	95.4	**100.0**	84.4	96.3	91.2
Modal - preterite	111	**100.0**	**99.1**	**100.0**	**98.2**	**98.2**	**100.0**	97.3	97.3	**99.1**	91.9	**100.0**	**99.1**	**100.0**	98.5
Modal - preterite subjunctive II	78	**88.5**	**89.7**	84.6	73.1	84.6	83.3	78.2	76.9	84.6	**89.7**	**88.5**	**89.7**	**93.6**	85.0
Modal negated - future I	82	98.8	98.8	100.0	100.0	100.0	100.0	100.0	97.6	98.8	100.0	100.0	100.0	98.8	99.4
Modal negated - future I subjunctive II	76	**100.0**	**100.0**	**100.0**	**98.7**	**98.7**	**100.0**	96.1	96.1	**100.0**	**100.0**	93.4	**100.0**	**98.7**	98.6
Modal negated - perfect	71	**98.6**	**98.6**	**98.6**	**98.6**	**100.0**	**97.2**	**97.2**	95.8	**100.0**	81.7	**100.0**	**98.6**	91.5	96.6
Modal negated - pluperfect	8	62.5	37.5	12.5	12.5	62.5	37.5	12.5	12.5	37.5	37.5	**100.0**	12.5	50.0	37.5
Modal negated - pluperfect subjunctive II	62	95.2	91.9	88.7	93.5	**100.0**	95.2	80.6	83.9	95.2	90.3	95.2	95.2	83.9	91.4
Modal negated - present	93	91.4	**98.9**	92.5	**100.0**	**98.9**	94.6	88.2	89.2	91.4	92.5	**100.0**	95.7	**100.0**	94.9
Modal negated - preterite	101	**100.0**	**100.0**	**100.0**	**99.0**	**100.0**	**98.0**	94.1	93.1	**99.0**	88.1	**99.0**	**99.0**	**100.0**	97.6
Modal negated - preterite subjunctive II	62	**98.4**	**98.4**	**98.4**	**100.0**	**100.0**	**100.0**	**98.4**	95.2	**98.4**	**100.0**	**100.0**	**98.4**	**96.8**	98.6
Progressive	19	**89.5**	**89.5**	**89.5**	**89.5**	**94.7**	**89.5**	**84.2**	**89.5**	**78.9**	47.4	68.4	**78.9**	26.3	78.1
Reflexive - future I	23	82.6	**100.0**	87.0	**100.0**	87.0	**100.0**	87.0	**91.3**	**100.0**	**95.7**	**100.0**	82.6	78.3	91.6
Reflexive - future I subjunctive II	25	80.0	**100.0**	80.0	**100.0**	88.0	**92.0**	88.0	**92.0**	**96.0**	80.0	**92.0**	**100.0**	72.0	89.2
Reflexive - future II	9	66.7	**88.9**	44.4	44.4	66.7	**88.9**	11.1	22.2	**100.0**	44.4	44.4	**100.0**	55.6	59.8
Reflexive - future II subjunctive II	10	**80.0**	**80.0**	**100.0**	**80.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	40.0	40.0	**100.0**	50.0	82.3
Reflexive - perfect	15	80.0	**93.3**	**86.7**	**100.0**	**100.0**	**93.3**	**86.7**	80.0	**93.3**	**86.7**	**93.3**	**100.0**	66.7	89.2
Reflexive - pluperfect	20	**75.0**	**85.0**	70.0	**95.0**	70.0	**90.0**	60.0	60.0	**90.0**	70.0	**80.0**	**95.0**	70.0	77.7
Reflexive - pluperfect subjunctive II	24	66.7	**91.7**	83.3	**91.7**	83.3	**95.8**	79.2	87.5	87.5	58.3	66.7	**100.0**	62.5	81.1
Reflexive - present	23	**91.3**	**100.0**	**95.7**	**100.0**	**100.0**	**95.7**	82.6	**91.3**	**95.7**	60.9	82.6	**100.0**	73.9	90.0
Reflexive - preterite	19	**89.5**	84.2	**100.0**	**100.0**	78.9	**94.7**	78.9	78.9	**100.0**	63.2	**89.5**	**89.5**	47.4	84.2
Reflexive - preterite subjunctive II	18	**94.4**	**94.4**	**94.4**	**100.0**	**94.4**	**94.4**	**88.9**	83.3	**94.4**	66.7	**88.9**	**100.0**	50.0	88.0
Transitive - future I	39	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	92.3	**100.0**	**100.0**	**100.0**	99.4
Transitive - future I subjunctive II	34	100.0	97.1	100.0	97.1	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.5
Transitive - future II	29	**100.0**	**96.6**	**100.0**	86.2	62.1	**100.0**	44.8	41.4	**100.0**	89.7	**100.0**	**100.0**	82.8	84.9
Transitive - future II subjunctive II	17	94.1	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	88.2	100.0	94.1	98.2
Transitive - perfect	41	**97.6**	**100.0**	**97.6**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	92.7	**100.0**	**100.0**	87.8	98.1
Transitive - pluperfect	31	**100.0**	**96.8**	45.2	**96.8**	22.6	**100.0**	45.2	54.8	**93.5**	**93.5**	**100.0**	**100.0**	80.6	79.2
Transitive - pluperfect subjunctive II	26	100.0	100.0	96.2	100.0	100.0	100.0	96.2	96.2	100.0	100.0	100.0	100.0	100.0	99.1
Transitive - present	43	**97.7**	**97.7**	**100.0**	**100.0**	**100.0**	93.0	**100.0**	**100.0**	**100.0**	**97.7**	**100.0**	**100.0**	**100.0**	98.9
Transitive - preterite	31	87.1	90.3	**96.8**	**100.0**	**100.0**	87.1	90.3	90.3	**100.0**	71.0	**93.5**	90.3	87.1	91.1
Transitive - preterite subjunctive II	29	72.4	69.0	69.0	**93.1**	**89.7**	69.0	69.0	69.0	65.5	44.8	**86.2**	72.4	62.1	71.6
Verb valency	81	**84.0**	**84.0**	**85.2**	**88.9**	**85.2**	**84.0**	**85.2**	**87.7**	77.8	77.8	**82.7**	**81.5**	65.4	82.2
Case government	28	**96.4**	**89.3**	**92.9**	**89.3**	**96.4**	**89.3**	**89.3**	**96.4**	**89.3**	78.6	**85.7**	**85.7**	71.4	88.5

categ	count	Onl-W	Onl-A	Onl-B	ChatG	Onl-M	Onl-Y	NLLBM	NLLBG	Onl-G	LanBr	GTCOM	ZengH	AIRC	avg
Mediopassive voice	18	83.3	**94.4**	**88.9**	**100.0**	**88.9**	**94.4**	83.3	83.3	72.2	**94.4**	**94.4**	**88.9**	77.8	88.0
Passive voice	19	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	94.7	100.0	100.0	89.5	98.8
Resultative predicates	16	43.8	43.8	50.0	62.5	43.8	43.8	62.5	62.5	37.5	37.5	43.8	43.8	12.5	45.2
micro-average	3234	**92.9**	**93.0**	91.2	**92.5**	88.3	**92.5**	85.6	85.6	91.5	81.2	90.7	89.4	82.2	89.0
phen. macro-average	3234	**91.0**	**91.3**	89.5	**91.5**	87.0	**91.6**	84.6	84.9	90.3	78.1	86.8	86.5	77.0	86.9
categ. macro-average	3234	**92.6**	**92.3**	**91.8**	**91.6**	90.1	89.2	89.2	88.9	88.1	82.3	82.0	75.6	74.3	86.8

Table 11: Accuracies (%) of successful translations on the phenomenon level for German–English. Boldface indicates the significantly best-performing systems per row.

categ	count	ChatG	Onl-W	Onl-B	Onl-A	Onl-Y	NLLBG	Onl-G	NLLBM	Onl-M	ZengH	LanBr	AIRC	avg
Ambiguity	24	**95.8**	**95.8**	**91.7**	**87.5**	**83.3**	**83.3**	**87.5**	**83.3**	**87.5**	**91.7**	75.0	50.0	84.4
Lexical ambiguity	24	**95.8**	**95.8**	**91.7**	**87.5**	**83.3**	**83.3**	**87.5**	**83.3**	**87.5**	**91.7**	75.0	50.0	84.4
Coordination & ellipsis	74	**90.5**	78.4	**93.2**	**85.1**	**93.2**	70.3	**90.5**	67.6	82.4	74.3	71.6	63.5	80.1
Gapping	12	**100.0**	75.0	**100.0**	**100.0**	**100.0**	58.3	**100.0**	50.0	**91.7**	**91.7**	75.0	75.0	84.7
Pseudogapping	7	**100.0**	**85.7**	**100.0**	**100.0**	**71.4**	**85.7**	**85.7**	**85.7**	**85.7**	**100.0**	**71.4**	42.9	84.5
Right node raising	15	**100.0**	**93.3**	80.0	80.0	**86.7**	**86.7**	**86.7**	**86.7**	80.0	73.3	**86.7**	80.0	85.0
Sluicing	14	**100.0**	**100.0**	**92.9**	**85.7**	**92.9**	**92.9**	**85.7**	**85.7**	**92.9**	78.6	**92.9**	78.6	89.9
Stripping	17	58.8	47.1	**94.1**	70.6	**100.0**	41.2	**94.1**	41.2	70.6	41.2	41.2	47.1	62.3
VP-ellipsis	9	**100.0**	**77.8**	**100.0**	**88.9**	**100.0**	66.7	**88.9**	66.7	**77.8**	**88.9**	66.7	44.4	80.6
False friends	33	**93.9**	**93.9**	**97.0**	**93.9**	**93.9**	**97.0**	**90.9**	**97.0**	**93.9**	**93.9**	**93.9**	81.8	93.4
Function word	41	**97.6**	**97.6**	**97.6**	**97.6**	**97.6**	**97.6**	**97.6**	**97.6**	**97.6**	75.6	**97.6**	85.4	94.7
Focus particle	22	95.5	95.5	95.5	95.5	95.5	95.5	95.5	95.5	95.5	95.5	95.5	90.9	95.1
Question tag	19	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	52.6	**100.0**	78.9	94.3
LDD & interrogatives	131	**96.9**	**96.2**	**95.4**	**96.9**	**96.9**	**94.7**	**93.9**	**93.9**	**93.9**	88.5	**92.4**	84.7	93.7
Extraposition	14	**85.7**	**85.7**	**78.6**	**85.7**	**78.6**	**78.6**	**64.3**	**71.4**	**71.4**	**78.6**	**57.1**	42.9	73.2
Inversion	13	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	92.3	100.0	100.0	100.0	99.4
Multiple connectors	17	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Negative inversion	17	**94.1**	**94.1**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**94.1**	**100.0**	76.5	**100.0**	**94.1**	96.1
Pied-piping	11	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	90.9	100.0	90.9	98.5
Polar question	8	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**87.5**	62.5	95.8
Preposition stranding	7	85.7	100.0	100.0	100.0	100.0	85.7	100.0	100.0	100.0	100.0	100.0	100.0	97.6
Split infinitive	11	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Topicalization	12	**100.0**	**83.3**	**83.3**	**91.7**	**91.7**	**91.7**	**83.3**	**91.7**	**83.3**	50.0	75.0	50.0	81.3
Wh-movement	21	100.0	100.0	95.2	95.2	100.0	90.5	95.2	90.5	95.2	95.2	100.0	95.2	96.0
Lexical Morphology	28	**85.7**	**85.7**	**82.1**	**75.0**	**67.9**	**67.9**	**64.3**	**64.3**	57.1	**82.1**	42.9	25.0	66.7
Functional shift	14	**92.9**	**85.7**	**78.6**	**85.7**	**64.3**	**85.7**	**71.4**	**71.4**	50.0	**78.6**	50.0	28.6	70.2
Noun formation (er)	14	**78.6**	**85.7**	**85.7**	**64.3**	**71.4**	**50.0**	**57.1**	**57.1**	**64.3**	**85.7**	35.7	21.4	63.1
MWE	95	**95.8**	**97.9**	**96.8**	91.6	**95.8**	85.3	89.5	86.3	86.3	92.6	78.9	68.4	88.8
Collocation	13	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**92.3**	**92.3**	**92.3**	**92.3**	**100.0**	**84.6**	69.2	93.6
Compound	17	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	94.1	99.5

categ	count	ChatG	Onl-W	Onl-B	Onl-A	Onl-Y	NLLBG	Onl-G	NLLBM	Onl-M	ZengH	LanBr	AIRC	avg
Idiom	12	**66.7**	**91.7**	**75.0**	50.0	**91.7**	41.7	33.3	41.7	25.0	50.0	16.7	0.0	48.6
Nominal MWE	20	**100.0**	**95.0**	**100.0**	**90.0**	85.0	85.0	**95.0**	**90.0**	85.0	**95.0**	70.0	70.0	88.3
Prepositional MWE	14	100.0	100.0	100.0	100.0	100.0	92.9	100.0	92.9	100.0	100.0	100.0	100.0	98.8
Verbal MWE	19	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**89.5**	**100.0**	**89.5**	**100.0**	**100.0**	**89.5**	63.2	94.3
Named entity & terminology	73	**95.9**	**95.9**	**95.9**	**97.3**	**97.3**	83.6	**94.5**	87.7	**94.5**	87.7	78.1	90.4	91.6
Date	13	**92.3**	**100.0**	**100.0**	**100.0**	**92.3**	**92.3**	**100.0**	**92.3**	**100.0**	69.2	61.5	**92.3**	91.0
Domainspecific Term	6	100.0	83.3	83.3	100.0	83.3	83.3	83.3	83.3	100.0	83.3	83.3	83.3	87.5
Location	17	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**94.1**	**100.0**	**94.1**	**100.0**	**100.0**	64.7	**100.0**	96.1
Measuring unit	18	**100.0**	**94.4**	**100.0**	**100.0**	**100.0**	66.7	**100.0**	77.8	83.3	**94.4**	83.3	**88.9**	90.7
Proper name	19	**89.5**	**94.7**	**89.5**	**89.5**	**100.0**	84.2	84.2	**89.5**	**94.7**	84.2	**94.7**	84.2	89.9
Negation	13	100.0	100.0	100.0	100.0	100.0	92.3	100.0	92.3	100.0	100.0	100.0	100.0	98.7
Non-verbal agreement	90	**97.8**	**94.4**	90.0	88.9	92.2	**94.4**	**93.3**	**95.6**	**95.6**	92.2	87.8	74.4	91.4
Coreference	29	**100.0**	**96.6**	86.2	86.2	89.7	**93.1**	89.7	**96.6**	**93.1**	89.7	79.3	58.6	88.2
Genitive	19	**100.0**	**94.7**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**89.5**	**100.0**	**94.7**	57.9	94.7
Personal Pronoun Coreference	12	**100.0**	**83.3**	58.3	58.3	66.7	**91.7**	75.0	**91.7**	**100.0**	66.7	66.7	**83.3**	78.5
Possession	27	92.6	96.3	100.0	96.3	100.0	96.3	100.0	96.3	100.0	100.0	100.0	96.3	97.8
Substitution	3	100.0	100.0	100.0	100.0	100.0	66.7	100.0	66.7	100.0	100.0	100.0	100.0	94.4
Punctuation	36	83.3	**97.2**	80.6	**88.9**	77.8	80.6	80.6	**86.1**	83.3	61.1	80.6	72.2	81.0
Quotation marks	36	83.3	**97.2**	80.6	**88.9**	77.8	80.6	80.6	**86.1**	83.3	61.1	80.6	72.2	81.0
Subordination	136	**99.3**	**97.1**	**97.8**	**97.8**	**96.3**	**97.8**	**97.8**	**97.8**	**97.8**	**97.1**	**99.3**	92.6	97.4
Adverbial clause	11	100.0	100.0	100.0	100.0	100.0	90.9	100.0	100.0	100.0	100.0	100.0	100.0	99.2
Cleft sentence	10	100.0	100.0	90.0	100.0	90.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	98.3
Contact clause	23	**100.0**	**95.7**	**95.7**	**100.0**	**100.0**	**100.0**	**95.7**	**100.0**	**95.7**	**100.0**	**100.0**	78.3	96.7
Indirect speech	12	91.7	83.3	100.0	91.7	91.7	100.0	100.0	100.0	91.7	91.7	100.0	100.0	95.1
Infinitive clause	10	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Object clause	8	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Pseudo-cleft sentence	14	100.0	100.0	100.0	100.0	92.9	100.0	92.9	100.0	100.0	100.0	100.0	92.9	98.2
Relative clause	34	**100.0**	**97.1**	**97.1**	**97.1**	**97.1**	**97.1**	**97.1**	**97.1**	**97.1**	91.2	**100.0**	**94.1**	96.8
Subject clause	14	100.0	100.0	100.0	92.9	92.9	92.9	100.0	85.7	100.0	100.0	92.9	85.7	95.2
Verb semantics	4	75.0	75.0	75.0	50.0	50.0	100.0	50.0	75.0	50.0	50.0	50.0	25.0	60.4
Verb tense/aspect/mood	2237	99.1	98.4	98.7	99.0	**99.6**	97.0	99.1	97.1	98.4	99.2	97.2	91.6	97.9
Conditional	19	94.7	89.5	94.7	94.7	94.7	94.7	94.7	94.7	89.5	94.7	84.2	78.9	91.7
Ditransitive - conditional I progressive	36	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**97.2**	91.7	99.1
Ditransitive - conditional I simple	42	**100.0**	**100.0**	**100.0**	**95.2**	**100.0**	73.8	**100.0**	76.2	92.9	**100.0**	73.8	66.7	89.9
Ditransitive - conditional II progressive	42	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	92.9	99.4
Ditransitive - conditional II simple	39	100.0	100.0	97.4	97.4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.6
Ditransitive - future I progressive	39	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	97.4	100.0	99.8
Ditransitive - future I simple	81	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	97.5	99.8
Ditransitive - future II progressive	7	**85.7**	**100.0**	**100.0**	**100.0**	**100.0**	**85.7**	**100.0**	**100.0**	**100.0**	**100.0**	**71.4**	0.0	86.9
Ditransitive - future II simple	21	**100.0**	**100.0**	**100.0**	**95.2**	**100.0**	**90.5**	**100.0**	85.7	**90.5**	**100.0**	71.4	9.5	86.9
Ditransitive - past perfect progressive	35	**100.0**	**97.1**	**94.3**	**100.0**	**100.0**	88.6	**100.0**	85.7	**94.3**	**100.0**	**100.0**	**97.1**	96.4
Ditransitive - past perfect simple	34	**97.1**	**94.1**	**97.1**	**100.0**	**100.0**	85.3	**100.0**	85.3	**94.1**	**100.0**	**100.0**	**100.0**	96.1
Ditransitive - past progressive	30	**96.7**	86.7	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	98.6

categ	count	ChatG	Onl-W	Onl-B	Onl-A	Onl-Y	NLLBG	Onl-G	NLLBM	Onl-M	ZengH	LanBr	AIRC	avg
Ditransitive - present perfect progressive	38	94.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	97.4	99.3
Ditransitive - present perfect simple	43	97.7	97.7	100.0	100.0	100.0	95.3	100.0	95.3	100.0	100.0	100.0	100.0	98.8
Ditransitive - present progressive	40	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	87.5	92.5	98.3
Ditransitive - simple past	48	100.0	100.0	100.0	100.0	100.0	95.8	100.0	97.9	100.0	100.0	100.0	95.8	99.1
Ditransitive - simple present	43	**100.0**	**97.7**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	90.7	90.7	98.3
Gerund	19	**94.7**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**94.7**	**100.0**	**94.7**	**94.7**	84.2	96.9
Imperative	9	88.9	100.0	88.9	100.0	100.0	100.0	100.0	88.9	100.0	77.8	100.0	77.8	93.5
Intransitive - conditional I progressive	24	100.0	95.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.7
Intransitive - conditional I simple	25	100.0	100.0	92.0	96.0	100.0	96.0	100.0	96.0	100.0	100.0	100.0	100.0	98.3
Intransitive - conditional II progressive	9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Intransitive - conditional II simple	20	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Intransitive - future I progressive	24	100.0	91.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0	95.8	100.0	100.0	99.0
Intransitive - future I simple	56	**100.0**	87.5	**98.2**	**100.0**	**100.0**	**98.2**	**98.2**	**98.2**	**98.2**	**100.0**	**100.0**	**100.0**	98.2
Intransitive - future II progressive	4	100.0	100.0	100.0	75.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	75.0	95.8
Intransitive - future II simple	24	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**95.8**	87.5	**100.0**	**100.0**	37.5	93.4
Intransitive - past perfect progressive	12	100.0	100.0	100.0	100.0	100.0	91.7	100.0	91.7	91.7	100.0	100.0	91.7	97.2
Intransitive - past perfect simple	25	100.0	96.0	100.0	100.0	100.0	96.0	100.0	96.0	100.0	100.0	100.0	100.0	99.0
Intransitive - past progressive	22	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Intransitive - present perfect progressive	4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Intransitive - present perfect simple	25	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	96.0	99.7
Intransitive - present progressive	49	**100.0**	93.9	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**98.0**	99.3
Intransitive - simple past	32	**100.0**	**100.0**	**100.0**	**96.9**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	84.4	98.4
Intransitive - simple present	33	97.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	93.9	99.2
Modal	283	98.9	**100.0**	98.9	**99.6**	**99.6**	98.6	**100.0**	98.9	**99.6**	**100.0**	98.9	**99.6**	99.4
Modal negated	251	**100.0**	**99.6**	98.8	**99.6**	**99.6**	98.8	**99.6**	98.8	**99.2**	**99.6**	**99.2**	98.4	99.3
Reflexive - conditional I progressive	23	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**95.7**	**100.0**	**100.0**	**100.0**	**95.7**	87.0	98.2
Reflexive - conditional I simple	22	100.0	100.0	100.0	90.9	100.0	90.9	100.0	90.9	95.5	95.5	95.5	90.9	95.8
Reflexive - conditional II progressive	12	100.0	91.7	100.0	100.0	100.0	100.0	91.7	100.0	100.0	100.0	91.7	100.0	97.9
Reflexive - conditional II simple	19	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	94.7	100.0	99.6
Reflexive - future I progressive	11	100.0	100.0	90.9	100.0	100.0	100.0	100.0	100.0	90.9	90.9	100.0	81.8	96.2
Reflexive - future I simple	33	100.0	100.0	100.0	100.0	100.0	100.0	97.0	100.0	100.0	100.0	100.0	100.0	99.7
Reflexive - future II progressive	5	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	20.0	93.3
Reflexive - future II simple	11	**100.0**	**100.0**	**81.8**	**100.0**	**100.0**	**81.8**	**100.0**	**90.9**	**100.0**	**100.0**	**100.0**	63.6	93.2
Reflexive - past perfect progressive	12	**100.0**	75.0	**91.7**	**100.0**	**100.0**	66.7	**100.0**	66.7	66.7	**100.0**	**100.0**	**91.7**	88.2
Reflexive - past perfect simple	22	**95.5**	**86.4**	**95.5**	**95.5**	**95.5**	72.7	**95.5**	72.7	**86.4**	**95.5**	**95.5**	**90.9**	89.8
Reflexive - past progressive	25	**100.0**	**100.0**	**100.0**	**100.0**	**96.0**	**100.0**	88.0	**100.0**	**100.0**	**100.0**	88.0	88.0	96.7
Reflexive - present perfect progressive	11	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Reflexive - present perfect simple	24	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	87.5	99.0
Reflexive - present progressive	20	100.0	100.0	90.0	95.0	90.0	95.0	90.0	100.0	95.0	95.0	95.0	90.0	94.6
Reflexive - simple past	25	**100.0**	**100.0**	**92.0**	**100.0**	**100.0**	**100.0**	**96.0**	**100.0**	**100.0**	**100.0**	**92.0**	88.0	97.3
Reflexive - simple present	14	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Transitive - future II progressive	5	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	0.0	91.7
Transitive - conditional I progressive	19	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	84.2	98.7

categ	count	ChatG	Onl-W	Onl-B	Onl-A	Onl-Y	NLLBG	Onl-G	NLLBM	Onl-M	ZengH	LanBr	AIRC	avg
Transitive - conditional I simple	12	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	75.0	**100.0**	75.0	**91.7**	**100.0**	**83.3**	**83.3**	92.4
Transitive - conditional II progressive	22	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	90.9	99.2
Transitive - conditional II simple	26	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	84.6	98.7
Transitive - future I progressive	23	**100.0**	**95.7**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	87.0	98.6
Transitive - future I simple	48	**100.0**	**100.0**	**100.0**	**97.9**	**97.9**	**100.0**	**100.0**	**100.0**	**97.9**	**97.9**	**100.0**	91.7	98.6
Transitive - future II simple	15	**93.3**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	6.7	91.7
Transitive - past perfect progressive	17	100.0	94.1	100.0	100.0	100.0	88.2	100.0	88.2	88.2	100.0	100.0	88.2	95.6
Transitive - past perfect simple	18	100.0	100.0	100.0	100.0	100.0	94.4	100.0	94.4	100.0	100.0	100.0	88.9	98.1
Transitive - past progressive	16	68.8	**93.8**	68.8	50.0	**100.0**	62.5	62.5	68.8	81.3	56.3	43.8	56.3	67.7
Transitive - present perfect progressive	20	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	90.0	99.2
Transitive - present perfect simple	29	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	93.1	99.4
Transitive - present progressive	28	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	96.4	92.9	99.1
Transitive - simple past	30	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**96.7**	**100.0**	**100.0**	**100.0**	**100.0**	90.0	98.9
Transitive - simple present	33	**100.0**	**100.0**	**97.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	90.9	99.0
Verb valency	94	**86.2**	**86.2**	**88.3**	**86.2**	**79.8**	77.7	76.6	**80.9**	**78.7**	**86.2**	72.3	59.6	79.9
Case government	20	90.0	90.0	95.0	95.0	90.0	90.0	90.0	95.0	90.0	90.0	85.0	75.0	89.6
Catenative verb	15	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**93.3**	80.0	97.8
Mediopassive voice	15	**80.0**	**73.3**	**73.3**	**66.7**	**53.3**	**46.7**	**40.0**	**60.0**	33.3	**73.3**	**40.0**	20.0	55.0
Passive voice	14	92.9	92.9	92.9	92.9	92.9	92.9	92.9	92.9	92.9	92.9	92.9	71.4	91.1
Resultative	16	87.5	87.5	87.5	93.8	87.5	75.0	87.5	81.3	93.8	93.8	75.0	68.8	84.9
Semantic roles	14	64.3	71.4	78.6	64.3	50.0	57.1	42.9	50.0	57.1	64.3	42.9	35.7	56.5
micro-average	3109	**97.8**	97.0	**97.2**	97.0	**97.4**	94.4	96.6	94.7	95.9	95.9	93.5	87.1	95.4
phen. macro-average	3109	**96.7**	95.6	**96.0**	95.3	95.8	91.9	94.5	92.0	93.7	93.6	90.3	80.3	93.0
categ. macro-average	3109	**92.9**	**92.6**	**92.0**	89.0	88.1	88.0	87.1	86.8	86.5	84.8	81.2	71.0	86.7

Table 12: Accuracies (%) of successful translations on the phenomenon level for English–German. Boldface indicates the significantly best-performing systems per row.

categ	count	Onl-G	Onl-W	Onl-B	ChatG	Onl-Y	Onl-A	NLLBM	NLLBG	Onl-M	PROMT	ZengH	LanBr	avg
Ambiguity	20	**70.0**	**60.0**	50.0	**85.0**	55.0	45.0	50.0	50.0	35.0	30.0	45.0	25.0	50.0
Lexical ambiguity	20	**70.0**	**60.0**	50.0	**85.0**	55.0	45.0	50.0	50.0	35.0	30.0	45.0	25.0	50.0
Coordination & ellipsis	89	**82.0**	**83.1**	67.4	**77.5**	68.5	65.2	62.9	66.3	67.4	58.4	50.6	49.4	66.6
Gapping	17	**88.2**	**76.5**	29.4	**64.7**	**76.5**	41.2	52.9	**64.7**	**70.6**	29.4	17.6	29.4	53.4
Pseudogapping	14	**78.6**	**78.6**	**57.1**	**64.3**	**35.7**	**42.9**	28.6	28.6	**42.9**	**50.0**	**50.0**	14.3	47.6
Right node raising	16	75.0	81.3	81.3	68.8	75.0	68.8	75.0	68.8	68.8	56.3	68.8	68.8	71.4
Sluicing	12	83.3	83.3	75.0	83.3	66.7	83.3	58.3	58.3	75.0	66.7	50.0	41.7	68.7
Stripping	16	**93.8**	**93.8**	81.3	**100.0**	68.8	**87.5**	**93.8**	**93.8**	**87.5**	**87.5**	75.0	75.0	86.5
VP-ellipsis	14	**71.4**	**85.7**	**85.7**	**85.7**	**85.7**	**71.4**	**64.3**	**78.6**	**57.1**	**64.3**	42.9	**64.3**	71.4
False friends	14	85.7	85.7	78.6	64.3	85.7	71.4	57.1	64.3	64.3	57.1	71.4	50.0	69.6
Function word	29	**96.6**	**96.6**	**96.6**	**93.1**	**82.8**	**82.8**	**96.6**	**93.1**	**96.6**	**86.2**	37.9	75.9	86.2
Focus particle	11	90.9	90.9	90.9	81.8	81.8	90.9	90.9	81.8	90.9	90.9	81.8	72.7	86.4

categ	count	Onl-G	Onl-W	Onl-B	ChatG	Onl-Y	Onl-A	NLLBM	NLLBG	Onl-M	PROMT	ZengH	LanBr	avg
Question tag	18	**100.0**	**100.0**	**100.0**	**100.0**	83.3	77.8	**100.0**	**100.0**	**100.0**	83.3	11.1	77.8	86.1
LDD & interrogatives	61	**95.1**	**95.1**	**91.8**	**88.5**	**93.4**	**91.8**	**88.5**	**88.5**	85.2	85.2	73.8	78.7	88.0
Inversion	13	**100.0**	**100.0**	**92.3**	**92.3**	**100.0**	**92.3**	**92.3**	**92.3**	76.9	**92.3**	**84.6**	**100.0**	92.9
Modifying Comparison	5	60.0	80.0	80.0	80.0	80.0	60.0	80.0	80.0	100.0	60.0	60.0	60.0	73.3
Multiple connectors	13	**100.0**	**100.0**	**92.3**	**92.3**	**92.3**	**100.0**	**92.3**	**92.3**	**84.6**	**92.3**	76.9	**84.6**	91.7
Pied-piping	7	100.0	100.0	100.0	100.0	85.7	85.7	100.0	100.0	85.7	100.0	85.7	85.7	94.0
Preposition stranding	9	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**88.9**	**88.9**	**88.9**	**88.9**	66.7	**88.9**	92.6
Topicalization	9	**100.0**	**77.8**	**88.9**	**88.9**	**88.9**	**88.9**	**88.9**	**77.8**	**77.8**	55.6	**88.9**	44.4	80.6
Wh-movement	5	80.0	100.0	80.0	40.0	100.0	100.0	60.0	80.0	100.0	100.0	20.0	60.0	76.7
Lexical Morphology	29	**86.2**	**86.2**	**75.9**	**86.2**	**65.5**	62.1	62.1	**65.5**	41.4	51.7	58.6	55.2	66.4
Functional shift	15	**86.7**	**100.0**	**93.3**	**93.3**	73.3	**86.7**	73.3	80.0	53.3	66.7	**86.7**	73.3	80.6
Noun formation (er)	14	**85.7**	**71.4**	**57.1**	**78.6**	**57.1**	35.7	**50.0**	**50.0**	28.6	35.7	28.6	35.7	51.2
MWE	71	**76.1**	**73.2**	**76.1**	**70.4**	59.2	**69.0**	**67.6**	**66.2**	**60.6**	**60.6**	**69.0**	54.9	66.9
Collocation	8	75.0	87.5	87.5	75.0	75.0	75.0	62.5	62.5	75.0	62.5	87.5	50.0	72.9
Compound	4	75.0	25.0	75.0	50.0	25.0	75.0	50.0	25.0	50.0	50.0	50.0	50.0	50.0
Idiom	14	50.0	50.0	50.0	57.1	21.4	35.7	21.4	28.6	14.3	28.6	28.6	28.6	34.5
Nominal MWE	17	88.2	88.2	88.2	82.4	82.4	94.1	88.2	88.2	88.2	82.4	82.4	76.5	85.8
Prepositional MWE	8	100.0	100.0	100.0	100.0	100.0	100.0	87.5	100.0	100.0	100.0	100.0	100.0	99.0
Verbal MWE	20	**75.0**	**70.0**	**70.0**	**60.0**	**50.0**	**55.0**	**80.0**	**70.0**	**50.0**	**50.0**	**70.0**	40.0	61.7
Named entity & terminology	71	**87.3**	**77.5**	**81.7**	73.2	69.0	**76.1**	63.4	63.4	69.0	59.2	**80.3**	60.6	71.7
Date	19	**94.7**	78.9	**100.0**	**89.5**	84.2	84.2	73.7	73.7	84.2	68.4	84.2	73.7	82.5
Domainspecific Term	9	77.8	66.7	77.8	55.6	55.6	66.7	22.2	22.2	33.3	44.4	77.8	33.3	52.8
Measuring unit	13	**92.3**	76.9	**92.3**	**92.3**	**84.6**	**84.6**	**92.3**	**100.0**	76.9	**92.3**	**92.3**	**92.3**	89.1
Onomatopeia	11	**72.7**	**90.9**	**63.6**	**54.5**	36.4	**63.6**	36.4	27.3	27.3	18.2	**81.8**	27.3	50.0
Proper name	6	100.0	66.7	66.7	66.7	66.7	66.7	83.3	83.3	100.0	66.7	66.7	66.7	75.0
Proper Name & Location	13	84.6	76.9	69.2	61.5	69.2	76.9	61.5	61.5	84.6	53.8	69.2	53.8	68.6
Negation	4	75.0	100.0	100.0	75.0	100.0	75.0	75.0	75.0	100.0	75.0	100.0	50.0	83.3
Non-verbal agreement	80	**76.3**	**86.3**	**75.0**	**82.5**	73.8	72.5	**81.3**	**81.3**	73.8	**75.0**	66.3	65.0	75.7
Coreference	23	**52.2**	**73.9**	**52.2**	**60.9**	**47.8**	**47.8**	**73.9**	**65.2**	**52.2**	**47.8**	39.1	34.8	54.0
Genitive	13	84.6	84.6	92.3	61.5	92.3	84.6	84.6	76.9	76.9	84.6	69.2	69.2	80.1
Personal Pronoun Coreference	17	82.4	**88.2**	70.6	**100.0**	64.7	76.5	**94.1**	**100.0**	**88.2**	82.4	76.5	82.4	83.8
Possessive Pronouns	16	**87.5**	**93.8**	**93.8**	**100.0**	**87.5**	**93.8**	81.3	81.3	81.3	**87.5**	**87.5**	**87.5**	88.5
Substitution	11	**90.9**	**100.0**	**81.8**	**100.0**	**100.0**	72.7	72.7	**90.9**	**81.8**	**90.9**	72.7	63.6	84.8
Punctuation	12	**100.0**	**83.3**	**91.7**	66.7	75.0	**100.0**	**83.3**	**83.3**	66.7	**91.7**	0.0	**91.7**	77.8
Quotation marks	12	**100.0**	**83.3**	**91.7**	66.7	75.0	**100.0**	**83.3**	**83.3**	66.7	**91.7**	0.0	**91.7**	77.8
Subordination	130	**93.8**	**96.9**	**93.8**	**93.8**	**93.8**	90.0	86.9	88.5	**93.8**	89.2	68.5	83.1	89.4
Adverbial clause	11	**81.8**	**100.0**	**81.8**	**100.0**	**81.8**	**81.8**	63.6	63.6	**81.8**	72.7	45.5	**90.9**	78.8
Cleft sentence	12	**100.0**	**100.0**	**91.7**	**91.7**	**91.7**	**91.7**	66.7	75.0	**91.7**	**91.7**	75.0	66.7	86.1
Complex object	18	94.4	94.4	94.4	94.4	88.9	94.4	88.9	94.4	94.4	94.4	77.8	77.8	90.7
Contact clause	10	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	90.0	80.0	97.5
Indirect speech	4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	75.0	100.0	100.0	75.0	100.0	95.8
Infinitive clause	21	**95.2**	**90.5**	**95.2**	**90.5**	**95.2**	**90.5**	**100.0**	**90.5**	**95.2**	**90.5**	66.7	**90.5**	90.9
Object clause	5	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	80.0	100.0	98.3

categ	count	Onl-G	Onl-W	Onl-B	ChatG	Onl-Y	Onl-A	NLLBM	NLLBG	Onl-M	PROMT	ZengH	LanBr	avg
Participle clause	20	85.0	95.0	90.0	90.0	90.0	85.0	80.0	85.0	95.0	85.0	85.0	80.0	87.1
Pseudo-cleft sentence	5	100.0	100.0	100.0	100.0	100.0	80.0	80.0	100.0	80.0	100.0	80.0	80.0	91.7
Relative clause	4	75.0	100.0	75.0	100.0	100.0	25.0	75.0	75.0	100.0	75.0	50.0	50.0	75.0
Subject clause	20	**100.0**	**100.0**	**100.0**	**90.0**	**100.0**	**100.0**	**95.0**	**100.0**	**95.0**	85.0	40.0	**90.0**	91.3
Verb semantics	17	**94.1**	**82.4**	**76.5**	47.1	58.8	**76.5**	52.9	47.1	58.8	58.8	58.8	41.2	62.7
Verb tense/aspect/mood	156	**91.7**	**94.2**	85.9	85.9	87.2	87.8	84.0	82.1	83.3	84.0	66.7	75.0	84.0
Conditional	24	**100.0**	**100.0**	**100.0**	**100.0**	**95.8**	**100.0**	**91.7**	87.5	87.5	**91.7**	45.8	87.5	90.6
Ditransitive	30	93.3	96.7	93.3	90.0	93.3	96.7	90.0	83.3	96.7	96.7	90.0	83.3	91.9
Gerund	15	**86.7**	**86.7**	**86.7**	66.7	**100.0**	80.0	73.3	53.3	80.0	73.3	53.3	53.3	74.4
Imperative	24	**87.5**	**95.8**	66.7	75.0	62.5	66.7	**83.3**	**95.8**	75.0	70.8	45.8	54.2	73.3
Intransitive	25	88.0	92.0	80.0	88.0	88.0	84.0	84.0	84.0	80.0	84.0	80.0	84.0	84.7
Reflexive	19	**89.5**	**89.5**	**78.9**	**84.2**	**89.5**	**89.5**	**68.4**	**68.4**	**73.7**	**78.9**	**78.9**	57.9	78.9
Transitive	19	**94.7**	**94.7**	**94.7**	**89.5**	**84.2**	**94.7**	**89.5**	**89.5**	**84.2**	**84.2**	63.2	**94.7**	88.2
Verb valency	126	**84.9**	**81.7**	**79.4**	**78.6**	**77.0**	72.2	68.3	64.3	73.0	70.6	69.8	60.3	73.3
Case government	25	96.0	96.0	92.0	96.0	92.0	88.0	84.0	84.0	96.0	92.0	84.0	84.0	90.3
Catenative verb	21	81.0	90.5	95.2	85.7	90.5	95.2	81.0	76.2	90.5	90.5	76.2	81.0	86.1
Impersonal Subject	5	100.0	100.0	100.0	80.0	100.0	100.0	100.0	60.0	100.0	100.0	100.0	80.0	93.3
Mediopassive voice	19	**84.2**	**63.2**	**73.7**	**73.7**	**57.9**	47.4	52.6	52.6	**68.4**	52.6	52.6	36.8	59.6
Passive voice	25	96.0	92.0	92.0	96.0	96.0	92.0	92.0	88.0	96.0	92.0	88.0	80.0	91.7
Resultative	16	**68.8**	**75.0**	**62.5**	**62.5**	**62.5**	**37.5**	25.0	18.8	18.8	**37.5**	**50.0**	18.8	44.8
Semantic roles	15	66.7	53.3	33.3	33.3	33.3	40.0	40.0	40.0	26.7	20.0	40.0	26.7	37.8
micro-average	909	**86.9**	**86.8**	81.7	81.7	78.3	78.0	75.2	74.8	75.4	72.9	65.0	65.7	76.9
phen. macro-average	909	**87.0**	**86.8**	82.7	81.2	79.1	78.0	74.7	73.9	76.3	73.5	65.9	65.5	77.0
categ. macro-average	909	**86.3**	**85.5**	81.3	77.9	76.3	75.8	72.0	71.9	71.3	68.9	61.1	61.1	74.1

Table 13: Accuracies (%) of successful translations on the phenomenon level for English–Russian. The boldface indicates the significantly best-performing systems per row.

IIIT HYD's Submission for WMT23 Test-suite task

Ananya Mukherjee and **Manish Shrivastava**
Machine Translation - Natural Language Processing Lab
Language Technologies Research Centre
International Institute of Information Technology - Hyderabad
ananya.mukherjee@research.iiit.ac.in
m.shrivastava@iiit.ac.in

Abstract

This paper summarizes the results of our test suite evaluation on 12 machine translation systems submitted at the Shared Task of the 8th Conference of Machine Translation (WMT23) for English-German (en-de) language pair. Our test suite covers five specific **domains** (*entertainment, environment, health, science, legal*) and spans five distinct **writing styles** (*descriptive, judgments, narrative, reporting, technical-writing*). We present our analysis through automatic evaluation methods, conducted with a focus on domain-specific and writing style-specific evaluations. Our test-suite is available at https://github.com/wmt-conference/wmt23-testsuites/tree/main/submissions/en-de/IIITHYD_TestSuite

1 Introduction

Neural Machine Translation has made significant strides and has achieved a level of quality that proves valuable in numerous everyday scenarios. Nonetheless, various assessment methods for Machine Translation suggest that there is still ample room for enhancement. One such evaluation approach, geared towards identifying translation deficiencies in a more systematic manner, involves the utilization of test suites or challenge sets. Unlike conventional evaluations that draw test sets from random everyday texts, test suites comprise sentences that are carefully curated or selected to assess the MT systems' competence in translating specific linguistic phenomena. In this context, we present the results obtained from applying these test suites, analyzing the performance of state-of-the-art systems concerning numerous linguistically-driven phenomena. These test suites were administered to 12 MT systems submitted during the 8th Conference of Machine Translation (WMT23) (Kocmi et al., 2023) for English–German language pair.

We have developed a comprehensive test suite that encompasses five distinct domains (entertainment, environment, health, science, legal) and spans five different writing styles (descriptive, judgments, narrative, reporting, technical writing). The primary objective of the test suite is not to gauge a system's overall translation performance, as this aspect is already evaluated through manual assessment and various additional metrics within the primary shared task. Instead, the test suite focuses on assessing the translational proficiency across diverse domains and writing styles.

2 Test suite details

Table 1 illustrates the distribution of sentences per domain and per writing style, with a total of 2268 sentences.

2.1 Sentence Selection

In order to ensure diversity and robustness in our test suite, we collected English sentences from a wide array of sources, including BBC NEWS, Children's Stories, Textbooks, Journals, and Legal Datasets. These sentences were then categorized into clusters based on several criteria, such as the count of Noun Phrases (NP), Verb Phrases (VP), Named Entities (NE), Subordinate Clauses (SC), Discourse Markers (DM), Punctuation (P), and Sentence Length (SL).

Within each domain, we chose to include 70% of the sentences from each cluster in our dataset, thereby augmenting the diversity and comprehensiveness of our test suite.

2.2 Evaluation

Our automatic evaluation process for the 12 systems is conducted in three phases. The first phase assesses the overall test suite, the second phase focuses on specific domains, and the third phase examines various writing styles. In addition to these automatic evaluations, we conducted manual

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 246–251
December 6–7, 2023.

Writing Style	Domain					Total
	Entertainment	*Environment*	*Health*	*Science*	*Legal*	
Descriptive		27	39	33		427
Judgements					348	449
Narrative		38	33	61		492
Reporting	427	374	399	458		552
Technical-writing		10	21			348
Total	99	348	132	1658	31	**2268**

Table 1: Test-suite statistics (Count of sentences in each domain per writing-style)

MT systems	COMETKIWI
ONLINE-B	0.847 (1)
ONLINE-Y	0.847 (1)
ONLINE-W	0.846 (3)
ONLINE-A	0.845 (4)
GPT4-5shot (Hendy et al., 2023)	0.842 (5)
ONLINE-G	0.841 (6)
ONLINE-M	0.839 (7)
Lan-BridgeMT (Wu and Hu, 2023)	0.833 (8)
NLLB_Greedy (NLLB Team et al., 2022)	0.831 (9)
NLLB_MBR_BLE	0.831 (9)
ZengHuiMT (Zeng, 2023)	0.815 (11)
AIRC (Rikters and Miwa, 2023)	0.809 (12)

Table 2: System-wise ranking based on COMETKIWI scores. Top five systems are highlighted in bold. Ranks are mentioned in brackets

analyses with the assistance of professional German speakers who aided us in identifying the errors made by the systems, providing valuable insights into their translation quality.

2.3 Experiment Setup

In this paper, we present the evaluation of 12 sysems with our test suite. The systems are part of the news translation task of the Eighth Conference on Machine Translation (WMT23). We cover the system outputs for English-German (en-de) language pair.

2.4 Automatic Evaluation

To evaluate the performance of the 12 submitted MT systems, we utilize COMETKIWI (Rei et al., 2022) scores, which offer quality estimation scores derived from the source sentence and MT output. Using these scores, we determine the system rankings, as outlined in Table 2. We chose COMETKIWI because it performed best among the other reference-free metrics in the recent WMT22 Metrics Shared Task (Freitag et al., 2022).

2.4.1 Domain-wise Evaluation

We have calculated COMETKIWI scores for each domain and presented them in Figure 1.

From this figure, we can deduce that ONLINE-B, ONLINE-Y, ONLINE-W, and ONLINE-A exhibit a high degree of consistency in their performance across all five domains.

However, it is worth noting that GPT4-5shot displayed subpar performance when applied to legal data, while NLLB_Greedy demonstrated comparatively lower performance in the context of environmental data.

Another important evident observation is that the machine translation (MT) systems exhibit a similar trend in both the health and science domains. This similarity may be attributed to the interconnected nature of these domains.

Notably, both ZengHuiMT and AIRC displayed consistently poor performance across all domains.

2.4.2 Writing-Style-wise Evaluation

We have computed COMETKIWI scores for sentence belonging to various writing styles and visualized the results in Figure 2.

ONLINE-W excels in narrative writing style sentences, but its performance declines significantly for technical writing style. In contrast, NLLB_Greedy performs poorly across descriptive, reporting, and technical writing styles.

Both ZengHuiMT and AIRC exhibit subpar performance across all the writing-styles. Additionally, GPT4-5Shot experiences a decline in its performance when it comes to judgments.

ONLINE-G, on the other hand, demonstrates better performance in technical writing and reporting styles.

Indeed, based on COMETKIWI scores, it is clear that both ONLINE-B and ONLINE-Y consistently outperformed other MT systems across a diverse array of writing styles and domains. This consistent

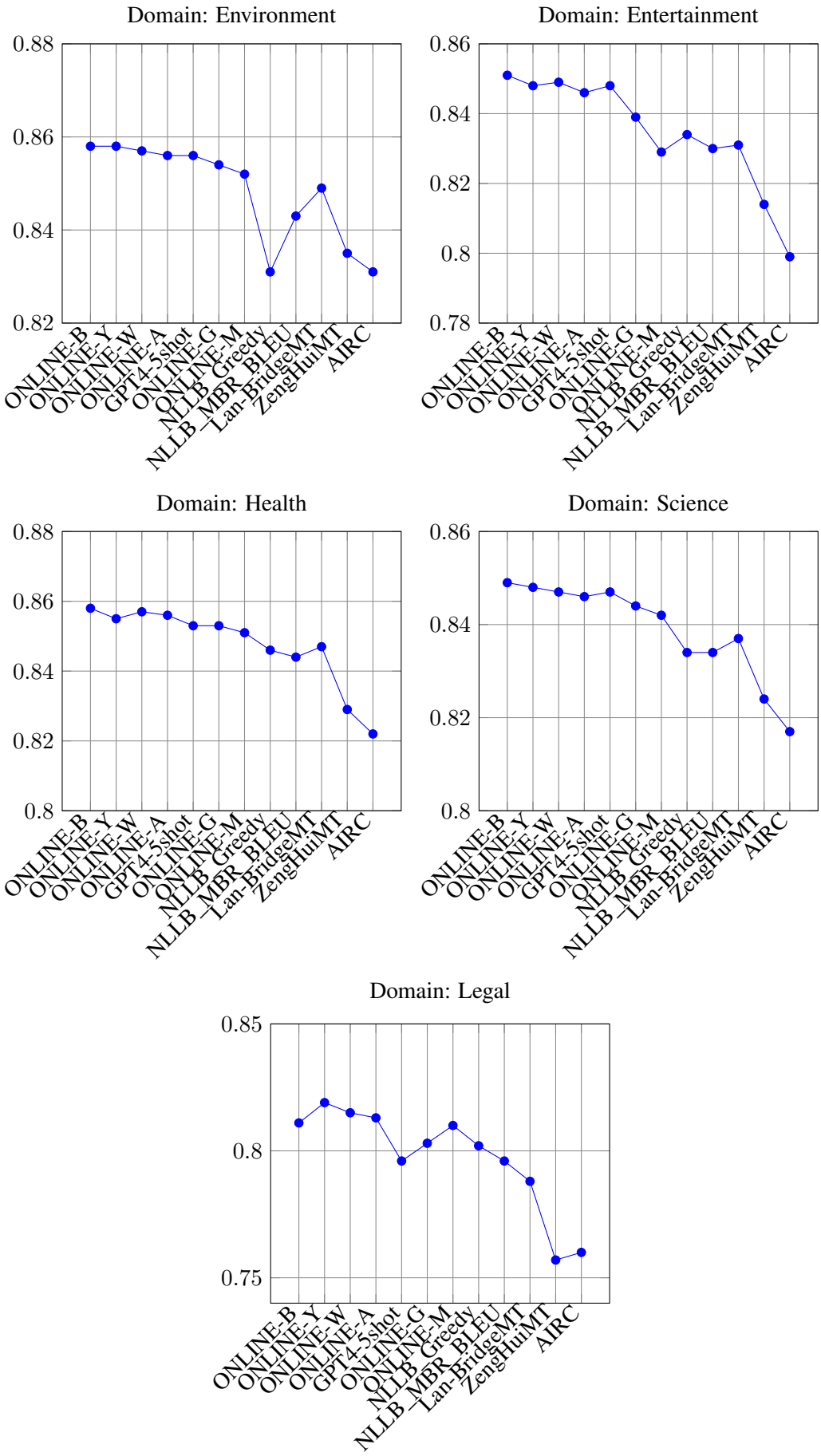


Figure 1: COMETKIWI scores of the systems with respect to domains

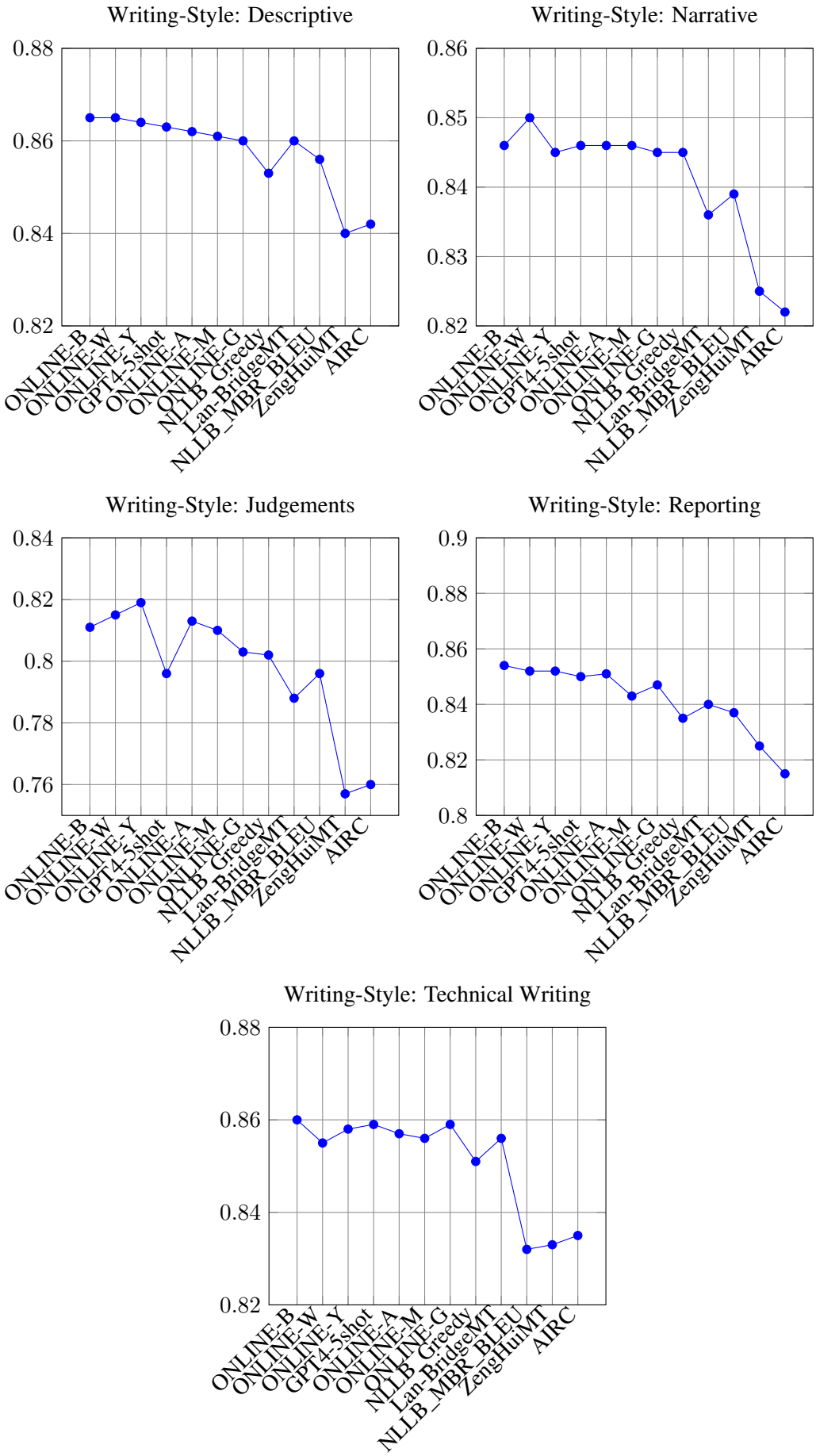


Figure 2: COMETKIWI scores of the systems with respect to writing-styles

superiority in performance suggests that these two MT systems are more robust and versatile, making them strong contenders for a wide range of translation tasks and scenarios.

2.5 Manual Assessments

These manual assessments are carried out voluntarily by professional German speakers who hold graduate-level qualifications and possess good knowledge in the domains covered by our test suite.

2.5.1 Gender-Neutral Pronouns

Machine translation (MT) systems often ascribe gender (sein/ihr $\sim$ his/her) to gender-neutral pronouns (it) in English. For instance, in the sentence '*Its age is not too dissimilar*,' ONLINE-B, ONLINE-M, ONLINE-G, ONLINE-A, ONLINE-W, Lan-BridgeMT, GPT4-5shot, and ZengHuiMT tended to assign the masculine gender 'Sein,' while the remaining systems ONLINE-Y, NLLB_Greedy, NLLB_MBR_BLEU, and AIRC preferred the feminine gender 'Ihr.' However, it's worth noting that in German, 'Sein' is typically used for neutral gender, thus introducing an intriguing linguistic nuance.

2.5.2 Repetition

Another intriguing factor is the phenomenon of Repetition, which is evident in cases like ZengHuiMT, where the translation includes additional information.

English source: a) Doing that amount is enough to reduce the risk of developing heart disease and stroke by 17% and cancer by 7%, the findings suggest.
b) While all living elements — the birds, animals and plants, forests, fisheries etc.— are biotic elements, abiotic elements include air, water, land etc.

Translation by ZengHuiMT: a) Die Ergebnisse deuten darauf hin, dass diese Menge ausreicht, um das Risiko für Herzerkrankungen und Schlaganfälle um 17 % und für Krebs um 7 % zu senken, so die Ergebnisse.
b) Während alle lebenden Elemente - Vögel, Tiere und Pflanzen, Wälder, Fischerei usw. - sind. Sie sind biotische Elemente, abiotische Elemente umfassen Luft, Wasser, Land usw.

Comment: a) The German translation is clear but includes an unnecessary repetition of **so die Ergebnisse** (the findings suggest) at the end.
b) Introduces an unnecessary repetition with **Sie sind biotische Elemente**.

2.5.3 Retention

Retention is another aspect that MT evaluation must consider. When it comes to challenging or complex words, retaining them might be permissible. However, for common or simpler words, retention should be heavily penalized.

Consider an example, "*These issues rarely have simple, single-discipline solutions that can be identified in one-off events or meetings.*" where ONLINE-B, ONLINE-M, GPT4-5shot, Lan-BridgeMT and AIRC MT systems retained the word *meetings* instead of translating it to *treffen*. This highlights the importance of addressing word retention in MT evaluation.

Manual assessments are indeed valuable for identifying gaps in machine translation quality. However, they come with significant drawbacks, including the need for extensive, non-reproducible human effort, time consumption, and high costs. Therefore, in addition to diverse test sets, it is crucial to develop robust automatic evaluation metrics capable of detecting and quantifying translation flaws efficiently and consistently.

3 Conclusion

This paper provides a comprehensive overview of our evaluation of 12 machine translation systems designed for the English-German language pair, all of which were submitted to the Shared Task during the 8th Conference on Machine Translation (WMT23). Our evaluation comprises a robust and diverse test-suite covering five distinct domains and encompassing five diverse writing styles. We conduct our analysis through a combination of automated assessments and manual evaluations, with a particular focus on domain-specific and writing style-specific performance. Based on our automatic evaluation, it is evident that both ONLINE-B and ONLINE-Y consistently surpassed other MT systems in performance across a diverse array of writing styles and domains.

References

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23). In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte M. Alves, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task.

Matīss Rikters and Makoto Miwa. 2023. Aist airc submissions to the wmt23 shared task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Yangjian Wu and Gang Hu. 2023. Exploring prompt engineering with gpt language models for document-level machine translation: Insights and findings. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Hui Zeng. 2023. Achieving state-of-the-art multilingual translation model with minimal data and parameters. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Test Suites Task: Evaluation of Gender Fairness in MT with MuST-SHE and INES

Beatrice Savoldi, Marco Gaido, Matteo Negri, Luisa Bentivogli
Fondazione Bruno Kessler
{bsavoldi,mgaido,negri,bentivo}@fbk.eu

Abstract

As part of the WMT-2023 "Test suites" shared task, in this paper we summarize the results of two test suites evaluations: MuST-SHE^{WMT23} and INES. By focusing on the en-de and de-en language pairs, we rely on these newly created test suites to investigate systems' ability to translate feminine and masculine gender and produce gender-inclusive translations. Furthermore we discuss metrics associated with our test suites and validate them by means of human evaluations. Our results indicate that systems achieve reasonable and comparable performance in correctly translating both feminine and masculine gender forms for naturalistic gender phenomena. Instead, the generation of inclusive language forms in translation emerges as a challenging task for all the evaluated MT models, indicating room for future improvements and research on the topic.

We make MuST-SHE^{WMT23} and INES freely available, respectively at:

https://mt.fbk.eu/must-she/

https://mt.fbk.eu/ines/

1 Introduction

As Machine Translation (MT) has made strides in generic performance, there is an increasing recognition of the need to scrutinize finer, more nuanced aspects that defy assessment through traditional metrics computed on generic test sets. It is within this context that the WMT Test Suites shared task emerges, aiming to provide a dedicated evaluation framework to delve into specific dimensions of MT output with a laser focus. In particular, those representing well-known challenges within the current MT landscape.

In light of the above, our contribution is dedicated to the critical themes of gender bias and inclusivity in translation (Savoldi et al., 2021). Given the large-scale deployment of MT, such aspects are not only relevant from a technical perspective, where gender-related errors negatively impact the accuracy of automatic translation. Rather, biased and non-inclusive systems can pose the concrete risk of under/misrepresenting gender minorities by over-producing masculine forms, while reinforcing binary gendered expectations and stereotypes (Blodgett et al., 2020; Lardelli and Gromann, 2022).

Accordingly, in this paper we present the FBK participation in the Test Suites shared task by conducting evaluations on two newly-created test suites:

1. **MuST-SHE**^{WMT23} for en-de, created as a English→German extension of the already existing multilingual MuST-SHE corpus (Bentivogli et al., 2020). This dataset is designed to allow for fine-grained analysis of (binary) gender bias in MT.

2. **INES** for de-en, designed to assess the ability of MT systems to generate inclusive language forms over non-inclusive ones when translating from German into English.

The MuST-SHE^{WMT23} and INES datasets, as well as their corresponding metrics and evaluations, are respectively discussed in Section 2 and 3. In the evaluations presented in this paper, we obtained translations of our test suites by systems that are part of the standard General Translation Task of the Eighth Conference on Machine Translation (WMT-2023). In particular, we evaluated 11 systems for MuST-SHE^{WMT23} en-de and 13 systems for INES de-en.

2 MuST-SHE^{WMT23}: en-de Evaluation

MuST-SHE^{WMT23} is a test suite designed to evaluate the ability of MT systems to correctly translate gender. It is composed of 200 segments that require the translation of at least one English gender-neutral word into the corresponding

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 252–262
December 6–7, 2023.

Form		Category 1: *Ambiguous first-person references*	Speaker
Fem.	SRC	The other hat that I've worn in my work is as **an activist**...	She
	REF_{De}	Der andere Hut, den ich bei meiner Arbeit getragen habe, ist **der**<den> **Aktivistin**<Aktivist>...	
Masc.	SRC	I mean, I'm a **journalist**.	He
	REF_{De}	Ich meine, ich bin **Journalist** <Journalistin>.	
		Category 2: *Unambiguous references disambiguated by gender info*	
Fem.	SRC	A college classmate wrote me a couple weeks ago and she said ...	He
	REF_{De}	**Eine**<Ein> **Kommilitonin**<Kommiliton> hat mir vor ein paar Wochen geschrieben und gesagt...	
Masc.	SRC	I decided to pay a visit to **the manager** [...] and he pointed ...	She
	REF_{De}	Also entschied ich mich **den**<die>**Filialleiter**<Filialleiterin> zu besuchen [...]	

Table 1: MuST-SHE annotated segments organized per category. For each gender-neutral word referring to a human entity in the English source sentence (SRC), the reference translation (REF) shows the corresponding gender-marked (Fem./Masc.) forms, annotated with their wrong <gender-swapped> forms. The last column of the table provides information about the speaker's gender.

masculine or feminine target word(s) in German. The test suite is created as an extension of MuST-SHE, a multilingual, natural benchmark built on TED talks data (Bentivogli et al., 2020), which allows for a fine-grained analysis of gender bias in MT and ST. The original MuST-SHE corpus comprises ∼3,000 (*audio, transcript, translation*) triplets annotated with qualitatively differentiated gender-related phenomena for thee language pairs: English→ French/Italian/Spanish. Here, we introduce a newly created **English→ German** textual portion (*transcript, translation*) of the MuST-SHE corpus.

2.1 MuST-SHE^{WMT23} Dataset

Phenomena of Interest. Following the MuST-SHE original design, MuST-SHE^{WMT23} is intended to evaluate the translation of a source English neutral word into its corresponding target gender-marked one(s) in the context of human referents, e.g. en: ***the*** *good* ***friend***, de: ***der/die*** *gute* ***Freund/in***.

To allow revealing a potential gap across the generation of feminine/masculine gender forms, the corpus includes a balanced number of feminine (F) and masculine (M) translation phenomena. Also, the corpus features two categories of phenomena, which differ in the presence/lack of a gender cue to disambiguate the translation. Namely, *i)* **CAT1**: consisting of first-person singular references (i.e. to the speaker), which are to be translated according to the speaker's linguistic expression of gender, e.g., *I am a good friend*. Then, *ii)* **CAT2** consisting of references to any participant, which are be translated according to explicit gender information available in the sentence, like lexically gendered words (*sister*, *Mr*), or pronouns (*He/she is a good friend*). These categories allow differentiating systems' behaviour across ambiguos vs. unambiguos cases.

Dataset creation. In order to create MuST-SHE^{WMT23} we collected a pool of English-German candidate segments by exploiting the same TED-based data sources used to create the other MuST-SHE datasets, namely: the Dev and Common Test sets of the MuST-C corpus, and other parallel sentences extracted from additional TED talks. Then, to target those segments that represented our phenomena of interest, we followed the same automatic procedure used for the original MuST-SHE benchmark, which was aimed to quantitatively and qualitatively maximize the extraction of an assorted variety of gender-marked phenomena. Regular expressions were employed to transform German gender-agreement rules into search patterns to be applied to our pool of candidate sentences. Also, to specifically match a differentiated range of gender-marked lexical items, we also compiled two series of 50 human-referring adjectives in English and German.

Once the automatic step was concluded, the pool of retrieved sentence pairs underwent a manual inspection to: *i)* remove any noise and keep only pairs containing at least one gender phenomenon; *ii)* ensure that the final (*transcript, translation*) pairs were not affected by misalignments resulting from the automatic procedure used to create

MuST-C and the new TED Talks data. Also, we examined the remaining pairs to verify that those to be included in MuST-SHE featured a a balanced distribution of categories, F/M forms, and speakers. Accordingly, since the MuST-C corpus presents a well-known gender imbalance[1], we excluded all of the extracted masculine segments that exceeded the feminine counterpart. Across categories, instead, we were not able to ensure a balanced distribution, as fewer instances from CAT1 could be identified.[2]

The resulting dataset – whose statistics are given in Table 2 – was then manually enriched with different types of information. For each segment, the annotation includes: category (1 and 2), gender form (F and M), and speaker's gender information.[3] Also, for each target gender-marked word in MuST-SHE^{WMT23}, we created a corresponding gender-swapped counterpart in the opposite gender form. As shown in Table 1, these word forms were paired and annotated in the reference translations. As we will describe in more detail in the upcoming Section 2.2, such annotated target gender-marked words are key features of MuST-SHE, which enable gender-sensitive, fine-grained analyses focusing solely on the correct generation of target gender-marked words.

The manual selection of appropriate sentences and their annotation was carried out by two annotators, both students proficient in the German language and with a background in Applied Linguistics.[4] Each annotator worked on half of the corpus independently and then revised the work done by the other. Finally, all the differences found were reconciled to get to the final corpus.

	CAT1	CAT2
Fem.	23	77
Masc.	23	77
Tot.	200	

Table 2: MuST-SHE^{WMT23} sentence-level statistics.

[1]As reported in MuST-Speakers, ∼70% of the speakers in MuST-C are referred to by *He* pronouns.

[2]This is most likely due to the gendered features of the German language, which – unlike es, fr, and it – does not carry gender markings on verbs (e.g., *I went* → de: *Ich bin gegangen* vs it: *Sono andat**a/o***) nor adjective in the nominative case (e.g., *I am good* → de: *Ich bin gut* vs. es: *Soy buen**a/o***.

[3]Such an information is migrated from the MuST-Speakers resource (Gaido et al., 2020), where gender information for each speaker in MuST-C has been labeled based on the personal pronouns the speakers used to describe themselves in their publicly available personal TED section.

[4]Their work was carried out as part of an internship at FBK.

2.2 MuST-SHE^{WMT23} Evaluation

Following the original MuST-SHE evaluation protocol described in Gaido et al. (2020), MuST-SHE^{WMT23} evaluation allows to focus on the gender realization of the target gender-marked forms, which are annotated in the reference translations together with their *wrong*, gender-swapped form (see Table 1). The evaluation is carried out in two steps, and by matching the annotated (*correct*/*wrong*) gender-marked words against the MT output. Accordingly, we first calculate the **Term Coverage** as the proportion of gender-marked words annotated in MuST-SHE (either in the correct or wrong form) that are actually generated by the system, on which the accuracy of gender realization is therefore *measurable*. Then, we define **Gender Accuracy** as the proportion of correct gender realizations among the words on which it is *measurable*. This evaluation method[5] has several advantages. On one side, *term coverage* unveils the precise amount of words on which systems' gender realization is measurable. On the other, *gender accuracy* directly informs about systems' performance on gender translation and related gender bias: scores below 50% indicate that the system produces the wrong gender more often than the correct one, thus signalling a particularly strong biased behaviour.

2.3 MuST-SHE^{WMT23} Results

In Table 3 we present the MuST-SHE^{WMT23} results for the 11 en-de systems that were submitted to the WMT-2023 standard General Translation Task. Starting from coverage results, the scores range between 67.34% (`AIRC`) and 77.07% (`ONLINE-G`), with only 3 systems under 70%. Hence, overall all models produce a good amount of gender-marked words that can be evaluated with regards to the accuracy of their gender realization. Moving onto the overall accuracy scores (All-Acc), we can see that – while there is still room for improvement – all of the evaluated MT systems are reasonably good at translating gender, with `ONLINE-M` emerging as the best model, able to correctly translate gender in 80% of the generated instances. If we go more fine-grained into results disaggregated across gender forms (F and M) and categories (1 and 2), however, we can unveil subtle differences. Indeed, for unambiguous

[5]The evaluation script is publicly available at: https://github.com/hlt-mt/FBK-fairseq/blob/master/examples/speech_to_text/scripts/gender/mustshe_gender_accuracy.py.

	All-Cov	All-Acc	1F-Acc	1M-Acc	2F-Acc	2M-Acc
ONLINE-M	75.07	80.07	50.00	84.00	86.08	80.00
ONLINE-Y	73.35	79.65	30.43	96.15	86.96	78.51
NLLB_MBR_BLEU	71.92	79.43	36.00	92.31	87.27	78.51
ONLINE-W	67.91	79.32	23.81	90.91	86.11	80.87
ONLINE-G	77.07	78.87	16.00	95.15	87.39	79.69
ONLINE-B	72.20	78.64	14.28	100.00	83.92	81.25
ONLINE-A	74.78	78.00	25.00	92.30	84.34	79.36
GPT4-5shot	69.63	77.94	10.53	95.83	83.33	80.17
ZenhHuiMT	73.07	77.35	19.23	95.65	84.82	79.37
Lan-BridgeMT	71.92	75.79	16.67	92.31	83.19	77.05
AIRC	67.34	73.98	10.53	87.50	81.25	74.56

Table 3: MuST-SHE^{WMT23} results for en-de. Systema are ranked based on overall Gender Accuracy (All-Acc).

gender translation from CAT2, systems perform basically on par across gender forms, with actually slightly higher results for feminine translation. Instead, results on CAT1 unveil a huge gender gap, with systems achieving almost perfect results for masculine translation, whereas feminine accuracy can be as low as 10.53%. In fact, the best ranked systems `ONLINE-M` generates the correct feminine form in 50% of the cases, namely at a random rate.

Overall, results on MuST-SHE^{WMT23} show that the evaluated MT systems are reasonably good at translating gender under realistic conditions, achieving comparable results across feminine and masculine gender translation. However, for ambiguous cases where the input sentence does not inform about the gender form to be used in translation, we confirm a strong skew where all systems favour masculine generation almost by default. This finding calls for further research endeavours and evaluation initiatives to counter gender bias in MT and measure future advances.

3 INES: de-en Evaluation

The INclusive Evaluation Suite (INES) is a test set designed to assess MT systems ability to produce gender-inclusive translations for the German→English language pair. By design, each German source sentence in INES includes an expression that can be rendered by means of either an *inclusive* (`IN`) or *non-inclusive* (`N-IN`) expression in the English target language.

Overall, INES comprises 162 manually curated German sentences, which are annotated with their corresponding (`IN`/`N-IN`) English expressions. As such, it allows to evaluate to what extent MT systems favor the generation of non-inclusive solutions over alternative, valid inclusive translation in their output.

3.1 INES Dataset

Here, we first describe the phenomena of interest included in INES. Then, we proceed by describing its creation methodology.

Phenomena of interest. Despite its comparatively restricted gender grammar, English has traditionally relied on the use of marked forms that treat the masculine gender as the conceptually generic, default human prototype, i.e. as *masculine generics* (Silveira, 1980; Bailey et al., 2022). Exemplary cases of such a phenomenon are man-derivates (e.g., *man-made*, *freshman*) and the use of masculine personal pronouns for generic referents (e.g., "each student must submit *his* form"). Besides, expressions such as "*man* and *wife*" have been identified as depicting skewed representation of genders and gender roles (Stahlberg et al., 2007). Toward the adoption of fairer language for all genders, alternative and inclusive solutions are increasingly promoted by institutions (Höglund and Flinkfeldt, 2023) and recommended in writing (APA, 2020). These include the use of unmarked forms (e.g. *human-made*, *first-year student*) and neutral pronouns (e.g. "each student must submit *their* form") for generic and under-specified referents, as well as more symmetrical formulations that cast men and women in the same role (e.g. "*husband* and *wife*").

On this basis, INES represents translation phenomena where, given a source German sentence, systems are confronted with the generation of a corresponding inclusive or non-inclusive solution. As shown by the examples in Table 4, the German sentences can entail the use of either *i)* a generic masculine form, e.g. *Der Polizist*, or *ii)* a term that does not convey gender, e.g. *Die Menschheit*. Regardless of the source German term, the expected ideal behaviour of the MT system always entails

	German src	English pair
a.	**Der Polizist** half der alten Dame, die Straße sicher zu überqueren.	**police officer**, **policeman**
b.	**Die Menschheit** hat das Potenzial, die Welt zu einem besseren Ort zu machen.	**humankind**, **mankind**
c.	Die **fachmännische** Arbeit des Teams führte zum erfolgreichen Abschluss des Projekts.	**skillful**, **workmanlike**
d.	Die **geschickte** Arbeit des Teams führte zum erfolgreichen Abschluss des Projekts.	**skillful**, **workmanlike**

Table 4: INES source German example sentences with their corresponding annotated English IN and N-IN terms.

the generation of inclusive target words.

Dataset Creation. Since the focus of the INES test suite is to evaluate the ability of MT systems to generate inclusive English translations, we started by compiling a list of well-established pairs of English IN/N-IN terms and expressions. This list was created based on existing collections of paired terms (Vanmassenhove et al., 2021; Amrhein et al., 2023) and integrated with few additional terms retrieved from other inclusive language guidelines from international institutions[6] and universities.[7][8] As a result, we obtained 48 IN/N-IN English pairs, which are shown in Table 5.

Starting from this list, we created the source German sentences that compose INES following a two-step semi-automatic procedure.

In the first step, for each English IN/N-IN term of the pairs, GPT[9] was prompted to generate 3 German sentences containing such term translated into German, for a total of 6 sentences for each English pair.

In the second step, the initial pool of 288 synthetic sentences was manually revised by a linguist proficient in German.[10] The revision was aimed to *i)* correct generation errors and *ii)* select a balanced amount of German sentences for each phenomenon of interest. To this purpose:

- when all the 6 German sentences generated for the two (IN/N-IN) terms of the English pair contained only gender-marked terms (e.g. *police officer –> Der Polizist / policeman –> Der Polizist*) or only gender-neutral terms (e.g. *humankind –> Die Menschheit / mankind –> Die Menschheit*), only 3 sentences out of 6 were kept (see examples a. and b. in Table 4);

IN vs N-IN for job titles	
anchor	anchorman
anchors	anchormen
bartender	barman
bartenders	barmen
business person	businessman
business persons	businessmen
chairpeople	chairmen
chairperson	chairman
firefighter	fireman
firefighters	firemen
flight attendant	steward
flight attendants	stewards
mail carrier	postman
mail carriers	postmen
member of congress	congressman
members of congress	congressmen
police officer	policeman
police officers	policemen
principal	headmaster
principals	headmasters
salesperson	salesman
salespersons	salesmen
spokesperson	spokesman
spokespeople	spokesmen
supervisor	foreman
supervisors	foremen
weather reporter	weatherman
weather reporters	weathermen
IN vs N-IN for generic man	
average person	average man
average people	average men
best people for the job	best men for the job
best person for the job	best man for the job
human-made	man-made
humankind	mankind
husband and wife	man and wife
intermediaries	middlemen
intermediary	middleman
skillful	workmanlike
laypeople	laymen
layperson	layman
workforce	manpower
first-year student	freshman
first-year students	freshmen
IN vs N-IN pronouns	
their	his
theirs	his
them	him
themself	himself
they	he

Table 5: INES pairs of English Inclusive (IN) vs non-inclusive (N-IN) expressions.

- on the contrary, when the 6 German sentences generated for the two (IN/N-IN) English terms included both gender-marked and gender-neutral forms (e.g. *firefighters –> Feuerwehrleute / firemen –> Feuerwehrmänner*), they were all kept, so as to have a richer representation of the phenomenon of interest

[6] https://www.europarl.europa.eu/cmsdata/151780/GNL_Guidelines_EN.pdf

[7] https://writingcenter.unc.edu/tips-and-tools/gender-inclusive-language/.

[8] https://www.gsws.pitt.edu/resources/faculty-resources/gender-inclusive-non-sexist-language-guidelines-and-resources.

[9] gpt-3.5-turbo.

[10] One of the authors of the paper.

in the source (see c. and d. in Table 4).

Unfortunately, we found only very few instances of double German realizations, and thus at the end of the manual revision, we remained with 162 German sentences: 21 with an inclusive source term, and 141 with a non-inclusive masculine generic in the source. All the 162 manually-curated German source sentences are included in INES, and provided with their corresponding English IN/N-IN term pair so as to allow for focused evaluations.

3.2 INES Evaluation

To evaluate systems against INES, we can leverage the annotated pairs of English IN/N-IN expressions and match them against the MT generated output. Accordingly, we can perform our evaluation by adopting the same evaluation protocol and metrics defined for MuST-SHE in 2.2. Namely, by *i)* first computing **Term Coverage** as the proportion of IN/N-IN generated by a system, and then *ii)* calculating **Inclusivity Accuracy** as the proportion of IN generated expressions, among all of the generated ones. As a result, all the *out of coverage words* (OOC) are necessarily left unevaluated.

While prior manual assessments of the terms left unevaluated by such an automatic method have been able to confirm the robustness and validity of the accuracy results in the context of binary gender translation (Savoldi et al., 2022b), here we hypothesise a potential limit for evaluating inclusivity in English outputs. Our hypothesis lies on the fact that English, a notional gender language (McConnell-Ginet, 2013), has a restricted repertoire of gender-marked – potentially N-IN – words, whereas most English nouns simply do not convey any gender distinctions (e.g., *doctor*, *secretary*, *president*). In other words, there might be many potential inclusive alternatives and synonyms (e.g. *presenter* and *host* for <*anchor*>) for a single N-IN term (e.g. <*anchorman*>). Thus, whereas OOC words in the context of binary gender present the same distribution assessed automatically in terms of accuracy, this metric might be stringent for inclusivity in English, and overly penalize the generation of alternative terms that differ from those annotated in INES.

In light of the above, we also propose the **Inclusivity Index** metric, defined as:

$$\text{Inclusivity Index} = 1 - \frac{n_{\text{N-IN}}}{N} \quad (1)$$

where $n_{\text{N-IN}}$ is the number of non-inclusive terms annotated in INES that are generated by a system, and N is the size of INES (i.e. total number of sentences to be evaluated).

In what follows, we thus carry out both **Inclusivity Accuracy** and **Inclusivity Index** evaluations,[11] and assess which one better correlates with human judgments.

3.3 INES Results

In this section (Table 6), we present the results obtained on INES by the 13 de-en systems that were submitted to the WMT-2023 standard General Translation Task. Such results are computed and discussed for Inclusivity Accuracy (Table 6a) and Inclusivity Index (Table 6b). Then, based on a manual analysis, we compare such automatic results against the systems ranking obtained with human evaluations (Table 6c).

Automatic Evaluation Results. Table 6a presents coverage and accuracy-based results. Based on such scores, the INES dataset emerges as quite a challenging test suite for current de-en systems. In fact, with the sole exception of the GPT4-5SHOT – which emerges as the best performing system (but see also Sec. 5) – all systems obtain scores that are below 50%, thus suggesting that they generate undesirable N-IN forms in more than half of the (measurable) cases. The lowest accuracy is obtained by NLLB_MBR_BLEU, amounting to 29.41% only.

Moving onto the Inclusivity Index results in Table 6b, from a bird's eye view we can immediately unveil some differences. On the one hand, GPT4-5SHOT and NLLB_MBR_BLEU still emerge as, respectively, the best and worst performing systems. On the other hand, however, there are discrepancies in the overall ranking. For instance, AIRC results as the system that generates the second-best level of inclusive translation according to the Inclusivity Index metrics, whereas it was ranked 7th in terms of accuracy.

Manual Evaluation Results. To verify which of the two automatic metrics yields more reliable results, we proceed with a manual analysis of all MT output sentences that defied the automatic evaluation procedure. Namely, we performed a human evaluation of all OOC terms to determine whether

[11] Evaluation script available at: https://github.com/hlt-mt/FBK-fairseq/blob/master/examples/speech_to_text/scripts/gender/INES_eval.py.

	Cov	Acc (↑)
GPT4-5shot	64.81	**65.71**
ONLINE-W	**75.31**	48.36
ONLINE-Y	74.07	45.83
ZenhHuiMT	73.46	44.54
ONLINE-A	74.69	42.98
ONLINE-B	70.99	41.74
AIRC	53.70	37.93
Lan-BridgeMT	68.52	36.94
ONLINE-M	70.37	36.84
ONLINE-G	74.07	35.00
GTCOM_Peter	74.69	33.06
NLLB_Greedy	74.07	31.67
NLLB_MBR_BLEU	73.46	29.41

(a) Coverage and Accuracy results

	In.Idx. (↑)
GPT4-5shot	77.78
AIRC	66.67
ONLINE-W	61.11
ONLINE-Y	59.88
ZenhHuiMT	59.26
ONLINE-B	58.64
ONLINE-A	57.41
Lan-BridgeMT	56.79
ONLINE-M	55.56
ONLINE-G	51.85
GTCOM_Peter	50.00
NLLB_Greedy	49.38
NLLB_MBR_BLEU	48.15

(b) Inclusivity Index results

	Human (↑)
GPT4-5shot	76.73
ONLINE-W	60.25
AIRC	59.03
ONLINE-Y	58.13
ZenhHuiMT	56.60
ONLINE-B	56.25
ONLINE-A	55.28
ONLINE-M	52.53
Lan-BridgeMT	52.26
ONLINE-G	48.45
NLLB_MBR_BLEU	46.25
GTCOM_Peter	48.13
NLLB_Greedy	44.03

(c) Human judgment – *Official ranking*

Table 6: INES evaluation results (percentage). Per each metric, systems are ranked based on their performance.

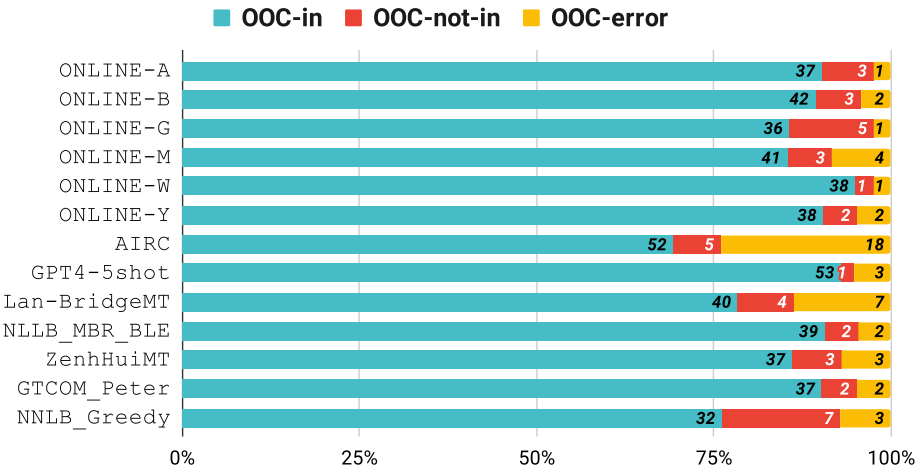


Figure 1: INES manual analysis results for out-of-coverage (OOC) terms.

Metric	Pearson (r)	Kendall (τ)	Spearman (ρ)
Acc	0.9601	0.8205	0.9285
In.Idx.	**0.9738**	**0.9231**	**0.9835**

Table 7: Correlation Coefficients with Human Judgment

the generated expression entailed *i)* an inclusive expression (OOC-in), which simply differed from the `IN` term annotated in INES but was completely acceptable; *ii)* a non-inclusive expression (OOC-not-in) different from the `N-IN` term annotated in INES; and finally *iii)* a translation error (OOC-error), which was not possible to judge in terms of inclusivity.[12] The results of such an analysis across all systems are reported in Figure 1. Such results show that, of all the OOC terms, the vast majority is represented by inclusive terms (e.g., <*business person*>/<*busissnessman*>→ *entrepeneur*). Errors, instead, are quite rare, just like non-inclusive OOC terms, which all correspond to the INES annotated `N-IN` term, but in a different number (e.g., <*freshmen*> → *freshman*).

In light of the above, our initial hypothesis – outlined in Sec. 3.2 – is thus reinforced: we do not find the same inclusivity distribution between evaluated cases in terms of accuracy (see Table 6a) and the OCC instances left unevaluated. Having now collected a complete evaluation of all the sentences, we leverage such information to obtain our official system ranking, which is shown in Table 6c. Results are computed as the proportion of inclusive (`IN` + OOC-in) terms generated by a system among all the terms that could be assessed (i.e. OOC-errors are not measurable, hence excluded).

[12]We underscore that such an analysis only concerns the terms representing the phenomena of our interest, whereas the overall judgement of the whole sentence is not accounted for.

Correlation between Automatic and Human evaluation. On this basis, and to finally verify our hypothesis, in Table 7 we report the correlation coefficients between the automatic metrics and human judgements. Accordingly, while both the Inclusivity Accuracy and Index show a satisfactory correlation with human judgements, the latter consistently emerges as a more reliable indicator of inclusivity. As such, Inclusivity Index is confirmed as the most suited measure to quantify gender-inclusive translation into English.

To conclude, our results in Tables 6 consistently indicate that current MT systems still struggle with the generation of inclusive translations. Within this landscape, GPT4-5SHOT consistently results as the model achieving the highest level of inclusivity, whereas all other models generate a ∼40% or more of non-inclusive translations in their output. This finding highlights that, while on the (binary) gender *bias* side (Section 2.3) MT systems still struggle with specific and particularly challenging ambiguous cases, the limitations of most of them on the gender *inclusion* side are evident and the problem emerges as an urgent topic for future research.

4 Related work

The last few years have witnessed and increasing attention toward (binary) gender bias in NLP (Sun et al., 2019; Stanczak and Augenstein, 2021; Savoldi et al., 2022a). Concurrently, emerging research has highlighted the importance of reshaping gender in NLP technologies in a more inclusive manner (Dev et al., 2021), also through the representation of non-binary identities in language (Lauscher et al., 2022; Ovalle et al., 2023). Foundational works in this area have included several applications, such as coreference resolution systems (Cao and Daumé III, 2020; Brandl et al., 2022) and fair rewriters (Vanmassenhove et al., 2021; Amrhein et al., 2023).

In MT, the research agenda has mainly focused on the improvement of masculine/feminine gender translation into grammatical gender languages (Savoldi et al., 2021). Along this line, different strategies have been devised to improve gender translation and mitigate masculine bias (Costa-jussà and de Jorge, 2020; Gaido et al., 2021; Choubey et al., 2021; Saunders et al., 2022). To test these methods and inspect systems' behaviour, several template-based datasets have been made available – such as WinoMT (Stanovsky et al., 2019) or SimpleGEN (Renduchintala and Williams, 2022) – which are especially intended to target occupational stereotyping. Instead, natural datasets such as the Arabic Parallel Gender Corpus (Alhafni et al., 2022) and GATE (Rarrick et al., 2023) allow for evaluation of gender bias under more naturalistic conditions. Among such type corpora, MuST-SHE (Bentivogli et al., 2020) represents the only multilingual, natural test set designed to evaluate gender bias for both MT and ST. Already available for English→French/Italian/Spanish, here we have contributed to its expansion for the English→German language pair.

As far as the topic of inclusivity and neutral language translation is concerned, research in MT is quite in its infancy. A notable exception is the work by Saunders et al. (2020), who created parallel test and fine-tuning data to develop MT systems able to generate non-binary translations for English→German/Spanish. However, their target sentences are artificial – created by replacing gendered morphemes and articles with synthetic placeholders – thus serving only as a proof-of-concept. Piergentili et al. (2023), instead, are the first to advocate for the use of target gender-neutral rephrasings and synonyms as a viable paradigm toward more inclusive MT when gender is unknown or simply irrelevant. Cho et al. (2019) and Ghosh and Caliskan (2023) investigate the preservation of gender-neutral pronouns for Korean/Bengali→English. Their results, however, show that current MT systems still face serious difficulties on relying on the inclusive, neutral pronoun *they* in translation. Along this line of work, INES – to the best of our knowledge – represents the first test suite designed to asses the use of neutral, inclusive forms beside pronouns for translating into English.

5 Conclusion

This paper summarizes the results of our WMT-2023 Test Suites evaluations, which focus on gender bias and inclusivity in translation. To this aim, we have introduced the en-de expansion of the multilingual MuST-SHE test set (Bentivogli et al., 2020) and the newly created INES dataset for de-en. The former is designed to assess gender bias and translation across a qualitatively differentiated selection of feminine/masculine gender phenomena. INES, instead, measures systems' ability to generate inclusive English translations that do not

rely on the use of masculine generics. Results on MuST-SHE^{WMT23} show that the evaluated MT systems are reasonably good at translating gender under realistic conditions, achieving comparable results across feminine and masculine gender translation. However, for ambiguous cases where the input sentence does not inform about the gender form to be used in translation, we confirm a strong skew where all systems tend to generate masculine forms almost by default. Results on INES, instead, indicate that providing inclusive translations still represents a quite challenging task for current MT systems, in spite of the increasingly widespread use and preference for inclusive language forms in English.

As a final remark, we acknowledge that the phenomena subject to our analysis (gender bias and gender inclusion) are not yet part of the repertoire of phenomena for which MT systems are currently designed. These systems are indeed primarily built with the goal of maximising translation quality in general rather than aspects of the problem, specifically fairness, for which sensitivity is still limited. All in all, however, this experience has allowed us to shed light on these issues, raise the awareness of the MT community and, hopefully, favour future developments.

Limitations

Naturally, this work comes with some limitations. First, both test suites are limited in size and number of language pairs considered. Despite their restricted size, however, both test suites provide a first glimpse into understanding and monitoring systems' behaviour with respect to gender and inclusivity. Additionally, rather than a limitation per se, both INES and MuST-SHE^{WMT23} are designed based on the specific linguistic features of the source and target language taken into account. As such, results in our evaluations intentionally do not aspire to scale and generalize to any language direction. Indeed, such linguistic specificity is also openly accounted for in the introduction of the new Inclusivity Index metric, which considers the morphology of English for a better-suited evaluation of gender inclusivity in MT. We also note that such a metric results as the best one for evaluating inclusivity under the given experimental conditions of this paper, where all the scrutinized systems (those submitted to the WMT General Translation task) are expected to feature generally good overall translation quality and to make few translation errors. As such, future work might be needed to further validate the stability of the Inclusivity Index metric under less optimal conditions and for different target languages, possibly proposing tailored metrics for each case. Finally, to generate the initial pool of sentences in INES we relied on the GPT (gpt-3.5-turbo) closed-source model. This has holds two types of implications. On the one hand, the use of proprietary models such as GPT has reproducibility consequences, since this model is regularly updated, thus potentially yielding future results that differ from those reported in this paper. On the other hand, relying on – even though only partially and post-edited – artificially generated data for testing models, might lead to contamination issues. Indeed, in Sec. 3.2 (Table 6) the GPT4-5SHOT model resulted as the most promising one, achieving the best results for inclusive translation. However, it remains to further verified whether our specific experimental settings and INES benchmark – where we use GPT-derived test data – have advantaged the performance of GPT4-5SHOT.

Ethics Statement

By addressing bias and inclusivity in MT, this work presents an inherent ethical component. It builds from concerns toward the societal impact of widespread translation technologies that reflect and propagate male-grounded and exclusionary language. Still, our work is not without risks either and thus warrants some ethical considerations. In particular, MuST-SHE^{WMT23} only focuses on traditional binary feminine/masculine gender forms. Also, INES investigates neutral, inclusive language in the context of generic, unknown referents and based on inclusive solutions encouraged by institutional guidelines. As such, we do not account for other non-binary solutions (e.g., neopronouns and neomorphemes) that are emerging from grassroots efforts. It should be stressed that the gendered and inclusive strategies incorporated in this MT work are not prescriptively intended. Rather, they are orthogonal to other attempts and non-binary expressions for inclusive language (technologies) (Lauscher et al., 2023; Ginel and Theroine, 2022).

Acknowledgements

This work is part of the project "Bias Mitigation and Gender Neutralization Techniques for Automatic Translation", which is financially supported

by an Amazon Research Award AWS AI grant. Also, we acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU. Also, we would like to thank the 2022 FBK internship students Sabrina Raus and Abess Benissmail from the University of Bolzano: the creation of MuST-SHE^{WMT23} was made possible by their work.

References

Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2022. User-centric gender rewriting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 618–631, Seattle, United States. Association for Computational Linguistics.

Chantal Amrhein, Florian Schottmann, Rico Sennrich, and Samuel Läubli. 2023. Exploiting biased models to de-bias text: A gender-fair rewriting model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4486–4506, Toronto, Canada. Association for Computational Linguistics.

APA. 2020. *Publication Manual of the American Psychological Association*, 7th edition. American Psychological Association.

April H Bailey, Adina Williams, and Andrei Cimpian. 2022. Based on billions of words on the internet, people= men. *Science Advances*, 8(13):eabm2463.

Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. Gender in Danger? Evaluating Speech Translation Technology on the MuST-SHE Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6923–6933, Online. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Stephanie Brandl, Ruixiang Cui, and Anders Søgaard. 2022. How conservative are language models? adapting to the introduction of gender-neutral pronouns. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3624–3630, Seattle, United States. Association for Computational Linguistics.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On Measuring Gender bias in Translation of Gender-neutral Pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, IT. Association for Computational Linguistics.

Prafulla Kumar Choubey, Anna Currey, Prashant Mathur, and Georgiana Dinu. 2021. Improving gender translation accuracy with filtered self-training. *arXiv preprint arXiv:2104.07695*.

Marta R. Costa-jussà and Adrià de Jorge. 2020. Fine-tuning neural machine translation on gender-balanced datasets. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 26–34, Barcelona, Spain (Online). Association for Computational Linguistics.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marco Gaido, Beatrice Savoldi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2020. Breeding Gender-aware Direct Speech Translation Systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3951–3964, Online. International Committee on Computational Linguistics.

Marco Gaido, Beatrice Savoldi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. How to split: the effect of word segmentation on gender bias in speech translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3576–3589, Online. Association for Computational Linguistics.

Sourojit Ghosh and Aylin Caliskan. 2023. Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages.

María Isabel Rivas Ginel and Sarah Theroine. 2022. Neutralising for equality: All-inclusive games machine translation. In *Proceedings of New Trends in Translation and Technology*, pages 125–133. NeTTT.

Frida Höglund and Marie Flinkfeldt. 2023. Degendering parents: Gender inclusion and standardised language in screen-level bureaucracy. *International Journal of Social Welfare*.

Manuel Lardelli and Dagmar Gromann. 2022. Gender-fair (machine) translation. In *Proceedings of New Trends in Translation and Technology*, pages 166–177. NeTTT.

Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022. Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1221–1232, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Anne Lauscher, Debora Nozza, Archie Crowley, Ehm Miltersen, and Dirk Hovy. 2023. What about em? how commercial machine translation fails to handle (neo-)pronouns.

Sally McConnell-Ginet. 2013. Gender and its Relation to Sex: The Myth of 'Natural' Gender. In Greville G. Corbett, editor, *The Expression of Gender*, pages 3–38. De Gruyter Mouton, Berlin, DE.

Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. "i'm fully who i am": Towards centering transgender and non-binary voices to measure biases in open language generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1246–1266.

Andrea Piergentili, Dennis Fucci, Beatrice Savoldi, Luisa Bentivogli, and Matteo Negri. 2023. Gender neutralization for an inclusive machine translation: from theoretical foundations to open challenges. In *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*, pages 71–83, Tampere, Finland. European Association for Machine Translation.

Spencer Rarrick, Ranjita Naik, Varun Mathur, Sundar Poudel, and Vishal Chowdhary. 2023. Gate: A challenge set for gender-ambiguous translation examples.

Adithya Renduchintala and Adina Williams. 2022. Investigating failures of automatic translationin the case of unambiguous gender. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3454–3469, Dublin, Ireland. Association for Computational Linguistics.

Danielle Saunders, Rosie Sallis, and Bill Byrne. 2020. Neural machine translation doesn't translate gender coreference right unless you make it. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 35–43, Barcelona, Spain (Online). Association for Computational Linguistics.

Danielle Saunders, Rosie Sallis, and Bill Byrne. 2022. First the worst: Finding better gender translations during beam search. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3814–3823, Dublin, Ireland. Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2022a. On the dynamics of gender learning in speech translation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 94–111, Seattle, Washington. Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2022b. Under the morphosyntactic lens: A multifaceted evaluation of gender bias in speech translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1807–1824, Dublin, Ireland. Association for Computational Linguistics.

Jeanette Silveira. 1980. Generic Masculine Words and Thinking. *Women's Studies International Quarterly*, 3(2-3):165–178.

Dagmar Stahlberg, Friederike Braun, Lisa Irmen, and Sabine Sczesny. 2007. Representation of the Sexes in Language. *Social communication*, pages 163–187.

Karolina Stanczak and Isabelle Augenstein. 2021. A Survey on Gender Bias in Natural Language Processing. *arXiv preprint arXiv:2112.14168*.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, IT. Association for Computational Linguistics.

Eva Vanmassenhove, Chris Emmery, and Dimitar Shterionov. 2021. NeuTral Rewriter: A rule-based and neural approach to automatic rewriting into gender neutral alternatives. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8940–8948, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Biomedical Parallel Sentence Retrieval using Large Language Models

Sheema Firdous
Fatima Jinnah Women University, Pakistan
sheemafirdous400@gmail.com

Sadaf Abdul Rauf
Fatima Jinnah Women University, Pakistan
Univ. Paris-Saclay, CNRS, LIMSI France
sadaf.abdulrauf@gmail.com

Abstract

We have explored the effect of in domain knowledge during parallel sentence filtering from in domain corpora. Models built with sentences mined from in domain corpora without domain knowledge performed poorly, whereas model performance improved by more than 2.3 BLEU points on average with further domain centric filtering. We have used Large Language Models for selecting similar and domain aligned sentences. Our experiments show the importance of inclusion of domain knowledge in sentence selection methodologies even if the initial comparable corpora are in domain.

1 Introduction

This paper describes FJWU's submission to the biomedical translation task. This year the focus of our research was domain specific parallel corpus mining from Wikipedia using Large Language Models, we explored the potential of the mined sentences using two sentence selection schemes. Neural Machine Translation (NMT) (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2015; Cho et al., 2014) has witnessed great success over the years (Vaswani et al., 2017; Zhang and Zong, 2020). NMT systems train on parallel corpora to produce translations that capture language intricacies and context with enormous precision as compared to the previous counterpart Statistical Machine Translation (SMT) systems.

Machine translation in the biomedical domain is becoming increasingly important due to the critical nature of medical scientific texts. The majority of these texts are published in English, and the goal of Biomedical Machine Translation is to make them accessible in multiple languages. However, this is a complex undertaking due to the extensive nature of this field and the vast and diverse vocabulary it encompasses. This vocabulary includes specialized terms and non-lexical forms (such as dates and biomedical entities) that pose unique challenges. Consequently, the quality of machine translation output fluctuates depending on the availability of biomedical resources tailored to each target language.

Availability of parallel corpora in reasonable amounts has greatly enhanced the performance of NMT systems, especially for the high-resource languages (Bojar et al., 2018). However, its efficacy remains sub optimal for low-resource languages and domain-specific contexts (Zoph et al., 2016; Koehn and Knowles, 2017; Lample et al., 2018; Chu and Wang, 2020). Performance of NMT system degrades as soon as the application domain deviates from training domain. Domain adaptation (Freitag and Al-Onaizan, 2016), transfer learning (Zoph et al., 2016; Khan et al., 2018; Abdul Rauf et al., 2020), model fusion (Gulcehre et al., 2015), back translation (Sennrich et al., 2015; Ul Haq et al., 2020), fine-tuning (Dakwale and Monz, 2017; Huck et al., 2018), data augmentation (Fadaee et al., 2017), data selective training (Van Der Wees et al., 2017; Knowles and Koehn, 2018), decoding strategies (Park et al., 2020), zero-shot translation (Johnson et al., 2017) are some of the techniques used to address this issue. We will be focusing on domain adaptation using data augmentation and fine tuning.

For this years submission we explore the potential of Large-scale Language Models for extracting parallel sentences from Wikipedia[1]. French-English parallel articles are scraped as detailed in Section 4. For learning sentence embeddings of scraped bilingual data, rather than training encoders from scratch, we leverage the potential of LLM in parallel sentence extraction from our bilingual scraped articles. We used LEALLA-Large, a lightweight system developed by (Mao and Nakagawa, 2023) to compute the language-agnostic low-dimensional sentence embeddings for each

[1]An online multilingual encyclopedia https://en.wikipedia.org/wiki/Main_Page

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 263–270
December 6–7, 2023.

sentence in the English and French parallel articles. Potential parallel sentences are filtered based on the similarity scores. These sentence are then further domain filtered by comparing the closeness with Medline Titles embeddings computed using Transformers MiniLM. Our experiments show the importance of inclusion of domain knowledge in sentence selection methodologies even if the initial comparable corpora are in domain. Our main contributions include:

- Presenting a methodology for domain inclusion in sentence retrieval tasks by using capabilities of Large Language Models
- Highlighting the importance of inculcation of in domain knowledge in sentence retrieval tasks even when the data source is in domain
- Release of the mined parallel corpora to the research community[2]

The paper is structured as follows: Section 2 presents a brief overview of background and related work, Section 3,4 elaborates the data collection pipeline, Section 5 outline the NMT experiments and results, followed by the conclusion of this study.

2 Related Work

Recent work on parallel sentence extraction has focused on lightweight end-to-end word-level and sentence-level embedding methods (Guo et al., 2018; Artetxe and Schwenk, 2018; Yang et al., 2019a). These embedding-based approaches have gained success (Grégoire and Langlais, 2017; Bouamor and Sajjad, 2018; Schwenk, 2018) as these systems outperformed the large-distributed computationally intensive systems (Uszkoreit et al., 2010; Abdul-Rauf and Schwenk, 2009) used to mine parallel documents. Bilingual sentence embeddings, learned from dual-encoder models, have also been used effectively for parallel corpus mining (Guo et al., 2018). Cross-lingual embeddings encode bilingual texts into a single unified vector space allowing nearest-neighbor search can be used to find potential translation candidates. These embedding approaches produce noisy matches that require a re-scoring step in order to obtain a clean parallel sentence retrieval as addressed by (Yang et al., 2019a) who explored using a bi-directional dual encoder with additive margin softmax (Wang et al., 2018) which results in state-of-the-art performance for sentence filtering. Multilingual sentence embedding approaches (Artetxe and Schwenk, 2018; Chidambaram et al., 2018) also show promising results.

Since language-specific models often demand extensive amounts of labeled data for training and can be limited by their language-specific parameters, language-agnostic sentence embedding(Artetxe and Schwenk, 2019; Yang et al., 2019b; Reimers and Gurevych, 2020; Feng et al., 2020; Mao et al., 2022) align multiple languages in a shared embedding space, facilitating parallel sentence alignment that extracts parallel sentences for training translation systems. Among them, LaBSE (Feng et al., 2020) achieved state-of-the-art performance on various bi-text retrieval. The problem of inference speed and computation overhead of large language models was addressed by (Mao and Nakagawa, 2023) who proposed Learning Leight-Weight Language-agnostic Sentence Embeddings (LEALLA) with Knowledge Distillation (Kim and Rush, 2016). They reported significant reduction in computation overhead and inference speed by providing language-agnostic low-dimensional sentence embeddings. We also use LEALLA in the second phase of our pipeline for parallel sentence alignemnent.

3 Wikipedia as a potential resource for biomedical data

Our primary objective was to collect a comprehensive dataset from the biomedical domain, we explored Wikipedia's key biological categories and selected those having a substantial volume of articles. A brief overview of the selected subdomains is given below:

1. **Biodbs** [3] refers to biological databases and contains links of a variety of biological databases.
2. **Genome Reference Consortium** is an international collaboration dedicated to creating and maintaining the most accurate and up-to-date Human Genome [4] reference sequence.

[2] https://github.com/sabdul111/Biomedical-Parallel-Corpus

[3] https://en.wikipedia.org/wiki/List_of_biological_databases

[4] https://en.wikipedia.org/wiki/Human_genome

Domain	Scraped URLs		Scraped Articles		Parallel Articles	Unique Articles
	French	English	French	English		
Biodbs	39.4K	77.3K	39.3K	68.7K	39.3K	1.2K
Human Genome	25.9K	59.1K	25.9K	49K	25.9K	25.9K
Health BioMed	42.8K	122.5K	42.8K	92.5K	42.8K	14.7K
NCBI	64.2K	133.8K	64K	133.6K	64K	51.2K
Pubmed	62.9K	134.5K	62.9K	117.4K	62.9K	22.4K
Total	235.2K	527.2K	234.9K	461.2K	234.9K	115.4K

Table 1: Scraped Data per subdomain

3. **National Institute of Biomedical Imaging and Bio engineering** plays a central role in advancing biomedical engineering research and provides a wealth of data and resources in the domain of Health Biomedical Engineering [5].

4. **The National Center for Biotechnology Information (NCBI)** [6] is a U.S. government agency that provides an extensive collection of biomedical and genomic resources.

5. **PubMed** [7] is a widely used online database maintained by the National Library of Medicine (NLM) which provides access to a vast collection of biomedical literature.

4 Parallel Corpus Mining

This section presents an overview of our parallel data creation pipeline. Wikipedia has been extensively used as a data resource for corpus development (Chu et al., 2014; Tufiş et al., 2013; Stefanescu et al., 2012; Karimi et al., 2018; Aghaebrahimian, 2018; Schwenk et al., 2019). We also used Wikipedia's inter language links to mine potential parallel sentences by exploring the potential of Large language models for filtering the closet candidates. Our data preparation pipeline involves three main steps; 1) Domain specific web scraping, 2) Candidate sentence scoring and filtering and 3) Domain adapted filtering.

Parallel article scrapping To extract the bilingual data we used Wikipedia's **Interwiki**[8] (also known as inter language links) property (Adafre and De Rijke, 2006; Otero and López, 2010; Chu et al., 2014; Aghaebrahimian, 2018). English Wikipedia has consistently held the distinction of possessing the highest article count among all language editions of Wikipedia. As of August 2023, there are 6,696,071[9] articles in English containing over 4.3 billion words.

We maximized recall in our article selection procedure by choosing English as the base language since it provided wider coverage of topics. Thus, for each unique English article, the corresponding French article (if found) was scrapped. We named the scrapped articles using the title of the English version, distinguishing them with *.en* for English and *.fr* for French files. At this stage, we had to retrieve the parallel articles since many of the English articles did not have the corresponding French articles (see Table 1). For parallel article retrieval, we compiled a list of all French articles and used this list to retrieve parallel English articles which resulted in our parallel French-English articles. The subdomains (see section § 3) had many overlapping articles which were removed and unique articles from each subdomain were selected.
Table 1 shows the amount of URLs, articles, parallel articles and the corresponding unique articles. At this stage we have unique parallel articles from each subdomain.

Parallel sentence filtering We used a lightweight pre-trained large language model LEALLA-Large (Mao and Nakagawa, 2023) which computes sentence embedding of 256 dimensions by distilling knowledge from LaBSE (Feng et al., 2020). It can be used to mine potential parallel sentences by finding the nearest neighbour of each source sentence in the target side according to cosine similarity, and filtering those below a threshold.

[5] https://en.wikipedia.org/wiki/Biomedical_engineering#Hospital_and_medical_devices

[6] https://en.wikipedia.org/wiki/National_Center_for_Biotechnology_Information

[7] https://en.wikipedia.org/wiki/PubMed

[8] The Interwiki property links the articles across various language editions of Wikipedia.

[9] https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

Domain	Parallel Sentences		
	Threshold 90	Threshold 85	Threshold 80
Biodbs	1,188	3,240	4,944
Human Genome	25,975	19,849	62,499
Health BioMed	14,677	41,555	66,008
NCBI	65,591	198,692	328,621
Pubmed	16,853	46,273	72,741
Total	124,284	309,609	534,813

Table 2: Parallel Sentences from the unique articles based on similarity threshold computed using LEALLA.

Parallel Sentences	BioFiltered Parallel Sentences		
	Threshold 20	Threshold 10	Threshold 0
Threshold 90	3,602	16,861	47,964
Threshold 85	15,286	64,888	169,215
Threshold 80	23,727	101,845	275,063
Total	42,615	183,594	492,242

Table 3: Bio-Filtered: Parallel sentences from Table 2 selected based on their proximity with Medline titles using MiniLM.

LEALLA Embedding vector is computed for each sentence in the French and English article. Thus for each French(source) sentence we have N potential matching sentences, where N is the number of sentences in English(target) article. The dot-product is then used to compute the similarity between each source and N target candidate sentences. The top 10 candidate sentences are retrieved for each sentence. At this stage we have a sorted list of potential parallel sentences from each subdomain.

It is important to note that these are potential bio med domain sentences since these are mined from in-domain articles. We focus on both precision and recall at this stage. Our sentence retrieval is recall oriented, given that English articles were roughly double the French articles, thus using French sentence as prompt to retrieve the matching English sentences promised a wider search space. For final parallel corpus creations we selected the sentences on similarity threshold. We report three thresholds (thresholds 80, 85, and 90) to retrieve parallel sentences from the retrieved top-10 sentence pairs. We are working on lower threshold sentences. A higher threshold indicates a greater degree of parallelism between the sentences. Table 2 shows the number of parallel sentences retrieved using different thresholds for each subdomain. We call these *LLMfilter* sentences for reference.

In domain filtering We did a second level selection from the *LLMfilter* parallel sentences extracted in the previous step. Even though these sentences come from bio-medical articles and are in-domain but our hypothesis is that there will be many sentences that may categorize as general domain. Our second filter is to ensure collection of purely biomedical sentences. For this we select Medline titles (Jimeno Yepes et al., 2017) as biomedical representative dataset since titles contain the main domain terminologies. An embedding was generated for Medline Titles using sentence transformers `paraphrase-multilingual-MiniLM-L12-v2`[10] which was then used to remove the out-domain sentences, striving to retain an optimal amount of in-domain sentences (pertaining to the biomedical domain). Dot product of each sentence with the Medline titles embedding was used to compute the similarity score(ranging from -1 to 1). We selected thresholds 20, 10, and 0 which correspond to 0.2, 0.1, and 0.0 respectively in the similarity score. Table 3 shows the number of sentences per threshold, we call these *Biofilter* sentences for reference.

Post-processing involved the removal of exceptionally short sentences, special characters, and sentences in languages other than the intended source and target languages. Duplicated and identical sentences were also removed from both English and French sides.

5 Translation performance on retrieved sentences

We used Transformer base (Vaswani et al., 2017) architecture provided by Fairseq (Ott et al., 2019)

[10] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

Model Name	Fine-tuning	WMT20 testset *LLMfilter*	Model Name	*Biofilter*
B1	-	19.52		
S1	B1 =>t90	18.12	SB1	20.29
S2	B1 =>t85	18.41	SB2	20.29
S3	B1 =>t80	18.54	SB3	20.58
S4	B1 =>t90-t85-t80	**18.78**	SB4	**21.11**
B2	-	38.71		
L5	B2 =>t90	19.69	LB1	21.81
L6	B2 =>t85	20.57	LB2	21.88
L7	B2 =>t80	**20.62**	LB3	22.07
L8	B2 =>t90-t85-t80	20.36	LB4	**22.43**

Table 4: BLEU scores on fine tuned datasets. B1 and B2 denote the baselines. B1 is trained on the biomedical texts provided by the WMT'23 organizers, while B2 is a big model trained on general domain and biomed data.

as `transformer_iwslt_en_de`. The ReLU activation function was used in all encoder and decoder layers. We optimize with Adam (Kingma and Ba, 2015), set up with a maximum learning rate of 0.0005 and an inverse square root decay schedule, as well as 4000 warmup updates.

All corpora were segmented into subword units using Sentence Piece (Kudo and Richardson, 2018) with a vocabulary of 32K units. We share the decoder input and output embedding matrices. Models are trained with mixed precision and a batch size of 4096 tokens on a single GPU. Systems were trained until convergence based on the BLEU score on the development sets. Evaluation was performed using SacreBleu (Post, 2018). Scores are chosen based on the best score on the development set (Medline 18, 19), and the corresponding scores for that checkpoint are reported on Medline 20 test set.

For fine-tuned systems, the process starts with models trained to convergence, based on BLEU score on dev sets. Training then resumes using a selected portion of the training corpus using the same parameters and criterion as for the base systems.

Baseline We trained a smaller model B1 on the biomedical texts provided by the WMT'23 organizers: Edp, Medline abstracts and titles (Jimeno Yepes et al., 2017), Scielo (Neves et al., 2016) and the Ufal Medical corpus[11] consisting of Cesta, Ecdc, Emea (OpenSubtitles), PatTR Medical and Subtitles. We used a bigger model B2 by (Xu et al., 2021) trained on WMT14 general domain corpus and WMT and supplementary biomed data including B1 data.

[11] https://ufal.mff.cuni.cz/ufal_medical_corpus

5.1 Results and Discussion

Table 4 presents the results using the two data selection methods. *LLMfilter* column shows the BLEU scores on Medline 20 testset for sentences filtered based on the sentence similarity score, whereas *Biofilter* are the sentences which were selected from the *LLMfilter* based on their closeness with the Biomedical Medline titles. Both filters used LLMs for computing similarity as detailed in section 4.

B1 represents a smaller baseline model trained on all biomed data provided by WMT organizers having a BLEU score of 19.52. This was further fine-tuned using each threshold dataset i.e. threshold 90, 85, and 80 (represented by t90, t85, and t80 respectively in 4), and finally with a concatenation of the 3 thresholds. Concatenation refers to the union of t90, t85, and t80. We did this to upsample the higher quality corpora (i.e. t90) to analyze the impact on MT. Evidently, none of the *LLMfilter* sentences improved the initial bio med baseline. The *Biofilter* sentences on the other hand helped improve the scores even when a small amount is added e.g. for t90 and the scores improved consistently with the increase in the number of sentences with SB4 yielding an increase of 1.59 BLEU points from the baseline. For the larger baseline B2, though none of the filtering schemes help improve the initial high score but still the supremacy of *Biofilter* sentences over *LLMfilter* is evident.

Arguably, both *LLMfilter* and *Biofilter* contain in-domain sentences as these have been selected from biomedical articles. The models built using the same thresholds for the two schemes have a difference of more than 2 BLEU points on average

with *Biofilter* systems being superior. Our results demonstrate the importance of inculcation of in-domain knowledge in sentence retrieval tasks even if the data source is in-domain as there are many sentences that do not pertain specifically to the domain and affect the results of domain-centered translation.

6 Conclusion

In this study, we explored the potential of large language models for parallel sentence extraction from domain-adapted bilingual corpus extracted from Wikipedia. On our dataset, we experimented with two data selection schemes and assessed the NMT performance for the biomedical domain. Our findings demonstrate that merely web-mining from in-domain corpus is not sufficient to improve domain-specific NMT performance but there is also a need for further filtering out out-domain sentences to improve the domain-specific NMT systems. Leveraging large language models to extract in-domain parallel sentences resulted in improved NMT performance by outperforming the baseline with 2 BLEU points.

Acknowledgments

This study is funded by the National Research Program for Universities (NRPU) by Higher Education Commission of Pakistan (5469/Punjab/NRPU/R&D/HEC/2016).

References

Abdul Rauf, S., Rosales Núñez, J. C., Pham, M. Q., and Yvon, F. (2020). Limsi @ wmt 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 803–812, Online. Association for Computational Linguistics.

Abdul-Rauf, S. and Schwenk, H. (2009). On the use of comparable corpora to improve SMT performance. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 16–23, Athens, Greece. Association for Computational Linguistics.

Adafre, S. F. and De Rijke, M. (2006). Finding similar sentences across multiple languages in wikipedia. In *Proceedings of the Workshop on NEW TEXT Wikis and blogs and other dynamic text sources*.

Aghaebrahimian, A. (2018). Deep neural networks at the service of multilingual parallel sentence extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1372–1383, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Artetxe, M. and Schwenk, H. (2018). Margin-based parallel corpus mining with multilingual sentence embeddings. *arXiv preprint arXiv:1811.01136*.

Artetxe, M. and Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bojar, O., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., and Monz, C. (2018). Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Bouamor, H. and Sajjad, H. (2018). Parallel sentence extraction from comparable corpora using multilingual sentence embeddings. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC), Miyazaki, Japan*, pages 7–12.

Chidambaram, M., Yang, Y., Cer, D., Yuan, S., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Learning cross-lingual sentence representations via a multi-task dual-encoder model. *arXiv preprint arXiv:1810.12836*.

Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Chu, C., Nakazawa, T., and Kurohashi, S. (2014). Constructing a chinese—japanese parallel corpus from wikipedia. In *LREC*, pages 642–647.

Chu, C. and Wang, R. (2020). A survey of domain adaptation for machine translation. *Journal of information processing*, 28:413–426.

Dakwale, P. and Monz, C. (2017). Fine-tuning for neural machine translation with limited degradation across in-and out-of-domain data. In *Proceedings of Machine Translation Summit XVI: Research Track*, pages 156–169.

Fadaee, M., Bisazza, A., and Monz, C. (2017). Data augmentation for low-resource neural machine translation. *arXiv preprint arXiv:1705.00440*.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2020). Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Freitag, M. and Al-Onaizan, Y. (2016). Fast domain adaptation for neural machine translation. *arXiv preprint arXiv:1612.06897*.

Grégoire, F. and Langlais, P. (2017). A deep neural network approach to parallel sentence extraction. *arXiv preprint arXiv:1709.09783*.

Gulcehre, C., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H.-C., Bougares, F., Schwenk, H., and Bengio, Y. (2015). On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Guo, M., Shen, Q., Yang, Y., Ge, H., Cer, D., Abrego, G. H., Stevens, K., Constant, N., Sung, Y.-H., Strope, B., et al. (2018). Effective parallel corpus mining using bilingual sentence embeddings. *arXiv preprint arXiv:1807.11906*.

Huck, M., Stojanovski, D., Hangya, V., and Fraser, A. (2018). Lmu munich's neural machine translation systems at wmt 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 648–654.

Jimeno Yepes, A., Névéol, A., Neves, M., Verspoor, K., Bojar, O., Boyer, A., Grozea, C., Haddow, B., Kittner, M., Lichtblau, Y., Pecina, P., Roller, R., Rosa, R., Siu, A., Thomas, P., and Trescher, S. (2017). Findings of the WMT 2017 biomedical translation shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 234–247, Copenhagen, Denmark. Association for Computational Linguistics.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1700–1709.

Karimi, A., Ansari, E., and Sadeghi Bigham, B. (2018). Extracting an English-Persian parallel corpus from comparable corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Khan, A., Panda, S., Xu, J., and Flokas, L. (2018). Hunter nmt system for wmt18 biomedical translation task: Transfer learning in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 655–661.

Kim, Y. and Rush, A. M. (2016). Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947*.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Knowles, R. and Koehn, P. (2018). Context and copying in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3034–3041, Brussels, Belgium. Association for Computational Linguistics.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Kudo, T. and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Lample, G., Ott, M., Conneau, A., Denoyer, L., and Ranzato, M. (2018). Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755*.

Mao, Z., Chu, C., and Kurohashi, S. (2022). Ems: efficient and effective massively multilingual sentence representation learning. *arXiv preprint arXiv:2205.15744*.

Mao, Z. and Nakagawa, T. (2023). Lealla: Learning lightweight language-agnostic sentence embeddings with knowledge distillation. *arXiv preprint arXiv:2302.08387*.

Neves, M., Yepes, A. J., and Névéol, A. (2016). The Scielo Corpus: a parallel corpus of scientific publications for biomedicine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2942–2948, Portorož, Slovenia. European Language Resources Association (ELRA).

Otero, P. G. and López, I. G. (2010). Wikipedia as multilingual source of comparable corpora. In *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC*, pages 21–25.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

Park, C., Yang, Y., Park, K., and Lim, H. (2020). Decoding strategies for improving low-resource machine translation. *Electronics*, 9(10):1562.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Reimers, N. and Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

Schwenk, H. (2018). Filtering and mining parallel data in a joint multilingual space. *arXiv preprint arXiv:1805.09822*.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., and Guzmán, F. (2019). Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*.

Sennrich, R., Haddow, B., and Birch, A. (2015). Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Stefanescu, D., Ion, R., and Hunsicker, S. (2012). Hybrid parallel sentence mining from comparable corpora. In *Proceedings of the 16th annual conference of the European association for machine translation*, pages 137–144.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Tufiş, D., Ion, R., Dumitrescu, Ş. D., and Stefanescu, D. (2013). Wikipedia as an smt training corpus. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 702–709.

Ul Haq, S., Abdul Rauf, S., Shaukat, A., and Saeed, A. (2020). Document level NMT of low-resource languages with backtranslation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 442–446, Online. Association for Computational Linguistics.

Uszkoreit, J., Ponte, J., Popat, A., and Dubiner, M. (2010). Large scale parallel document mining for machine translation.

Van Der Wees, M., Bisazza, A., and Monz, C. (2017). Dynamic data selection for neural machine translation. *arXiv preprint arXiv:1708.00712*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, F., Cheng, J., Liu, W., and Liu, H. (2018). Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930.

Xu, J., Pham, M. Q., Abdul Rauf, S., and Yvon, F. (2021). LISN @ WMT 2021. In *Proceedings of the Sixth Conference on Machine Translation*, pages 232–242, Online. Association for Computational Linguistics.

Yang, Y., Abrego, G. H., Yuan, S., Guo, M., Shen, Q., Cer, D., Sung, Y.-H., Strope, B., and Kurzweil, R. (2019a). Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. *arXiv preprint arXiv:1902.08564*.

Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Abrego, G. H., Yuan, S., Tar, C., Sung, Y.-H., et al. (2019b). Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307*.

Zhang, J. and Zong, C. (2020). Neural machine translation: Challenges, progress and future. *Science China Technological Sciences*, 63(10):2028–2050.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201*.

The Path to Continuous Domain Adaptation Improvements by HW-TSC for the WMT23 Biomedical Translation Shared Task

Zhanglin Wu, Daimeng Wei, Zongyao Li, Zhengzhe Yu, Shaojun Li, Xiaoyu Chen, Hengchao Shang, Jiaxin Guo, Yuhao Xie, Lizhi Lei, Hao Yang, Yanfei Jiang
Huawei Translation Service Center, Beijing, China
{wuzhanglin2,weidaimeng,lizongyao,yuzhengzhe,lishaojun18,chenxiaoyu35,
shanghengchao,guojiaxin1,xieyuhao2,leilizhi,yanghao30,jiangyanfei}@huawei.com

Abstract

This paper presents the domain adaptation methods adopted by Huawei Translation Service Center (HW-TSC) to train the neural machine translation (NMT) system on the English↔German (en↔de) WMT23 biomedical translation task. Our NMT system is built on deep Transformer with larger parameter sizes. Based on the biomedical NMT system trained last year, we leverage Curriculum Learning, Data Diversification, Forward translation, Back translation, and Transductive Ensemble Learning to further improve system performance. Overall, we believe our submission can achieve highly competitive result in the official final evaluation.

1 Introduction

Machine translation (MT) (Lopez, 2008) refers to the automatic translation of text from one language to another. The WMT23 biomedical translation task aims to evaluate the performance of MT systems in the biomedical domain. Due to the lack of sufficient in-domain data, domain adaptation (Chu and Wang, 2018; Wu et al., 2023) has naturally become the main research direction of this task.

This paper presents the domain adaptation methods adopted by HW-TSC to train the NMT (Bahdanau et al., 2015) system on en↔de language pair of the WMT23 biomedical translation task. Our method is mainly based on previous works (Wei et al., 2022, 2021; Yang et al., 2021). We try to train a domain classifier to select biomedical data from general data, then perform multi-step data cleaning on the selected in-domain data and keep only a high-quality subset for training. Based on the biomedical NMT system trained last year, we leverage Curriculum Learning (Zhang et al., 2019), Data Diversification (Nguyen et al., 2020), Forward Translation (Abdulmumin, 2021), Back Translation (Sennrich et al., 2016), and Transductive Ensemble Learning (Wang et al., 2020b) to further improve system performance.

Our system report includes four parts. Section 2 focuses on our data processing strategies while section 3 describes our training details. Section 4 explains our experiment settings and training processes, and section 5 presents the results.

2 Data

2.1 Data Volume

We obtain bilingual and monolingual data from various data sources, except medical database. Then, we use biomedical data and general data to train a domain classifier based on fasttext (Joulin et al., 2016) to select biomedical data from general data. Table 1 lists the final size of the training data.

language pairs	bitext data	monolingual data
en↔de	11.6M	en: 12.3M, de: 10.1M

Table 1: Bilingual and monolingual used for training.

2.2 Data Pre-processing

Our data processing procedure is basically the same as our method last year (Wu et al., 2022), including deduplication, XML content processing, langid (Lui and Baldwin, 2012) and fast-align (Dyer et al., 2013) filtering strategies, etc. As we use the same data pre-processing strategy as last year's, we will not go into details here.

2.3 Data Denoising

Since there may be some semantically dissimilar sentence pairs in bilingual data, we use LaBSE (Feng et al., 2022) to calculate the semantic similarity of each bilingual sentence pair, and exclude bilingual sentence pairs with a similarity score lower than 0.7 from the training corpus.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 271–274
December 6–7, 2023.

3 System Overview

3.1 Model

We continue using Transformer (Vaswani et al., 2017) as our neural machine translation (NMT) model architecture. As we did last year, we use a 25-6 deep model architecture. The parameters of the model are the same as Transformer-big. We just change the post-layer normalization to the pre-layer normalization, and set encoder layers to 25.

3.2 Curriculum Learning

A practical curriculum learning (CL) (Zhang et al., 2019) method should address two main questions: how to rank the training examples, and how to modify the sampling procedure based on this ranking. For ranking, we choose to estimate the difficulty of training samples according to their domain feature (Wang et al., 2020a). The calculation formula of domain feature is as follows, where θ_{in} represents an in-domain NMT model, and θ_{out} represents an out-of-domain NMT model.

$$q(x, y) = \frac{\log P(y|x; \theta_{in}) - \log P(y|x; \theta_{out})}{|y|} \quad (1)$$

For sampling, we adopt a probabilistic CL strategy[1] that takes advantage of the spirit of CL in a nondeterministic fashion without discarding the good practice of original standard training, like bucketing and mini-batching.

3.3 Data Diversification

Data Diversification (DD) (Nguyen et al., 2020) is a data augmentation method to boost NMT performance. It diversifies the training data by using the predictions of multiple forward and backward models and then merging them with the original dataset on which the final NMT model is trained. DD is applicable to all NMT models. It does not require extra monolingual data, nor does it add more computations or parameters. To conserve training resources, we only use one forward model and one backward model when performing DD.

3.4 Forward Translation

Forward translation (FT) (Abdulmumin, 2021), also known as self-training, is one of the most commonly used data augmentation methods. FT has proven effective for improving NMT performance by augmenting model training with synthetic parallel data. Generally, FT is performed in three steps: (1) randomly sample a subset from the large-scale source-side monolingual data; (2) use a "teacher" NMT model to translate the subset data into the target language to construct the synthetic parallel data; (3) combine the synthetic and authentic parallel data to train a "student" NMT model.

3.5 Back Translation

An effective method to improve NMT with target monolingual data is back translation (BT) (Sennrich et al., 2016; Wei et al., 2023). There are many works broaden the understanding of BT and investigates a number of methods to generate synthetic source sentences. Edunov et al. (2018) find that back translations obtained via sampling or noised beam outputs are more effective than back translations generated by beam or greedy search in most scenarios. Caswell et al. (2019) show that the main role of such noised beam outputs is not to diversify the source side, but simply to tell the model that the given source is synthetic. Therefore, they propose a simpler alternative strategy: Tagged BT. This method uses an extra token to mark back translated source sentences, which generally outperforms noised BT. For better joint use with FT, we use sampling back translation (ST).

3.6 Transductive Ensemble Learning

Ensemble learning (Garmash and Monz, 2016), which aggregates multiple diverse models for inference, is a common practice to improve the performance of machine learning models. However, it has been observed that the conventional ensemble methods only bring marginal improvement for NMT when individual models are strong or there are a large number of individual models. Transductive Ensemble Learning (TEL) (Zhang et al., 2019) studies how to effectively aggregate multiple NMT models under the transductive setting where the source sentences of the test set are known. TEL uses all individual models to translate the source test set into the target language space and then fine-tune a strong model on the translated synthetic data, which significantly boosts strong individual models and benefits a lot from more individual models.

[1] https://github.com/kevinduh/sockeye-recipes/tree/master/egs/curriculum

4 Experiment Settings

We use the open-source fairseq (Ott et al., 2019) for training, then we use SacreBLEU (Post, 2018) and multi-eval tool [2] to measure system performances. The main parameters are as follows: each model is trained using 8 A100 GPUs, batch size is 6144, parameter update frequency is 1, and learning rate is 5e-4. The number of warmup steps is 4000, and model is saved every 1000 steps. The architecture we used is described in section 3.1. We adopt dropout (Srivastava et al., 2014), and the rate varies across different training phases. When the training data is higher than tens of millions, the dropout ratio is set to 0.1, otherwise it is set to 0.3.

5 Results

Regarding en↔de, we use Curriculum Learning (CL), Data Diversification (DD), Forward Translation (ft), Back Translation (BT), and Transductive Ensemble Learning (TEL). The evaluation results of en→de and de→en NMT system on WMT22 biomedical test set are shown in Tables 2.

We see that CL can stably bring 3 SacreBLEU and multi-eval improvement, while DD, FT & ST and TEL can further slightly improve SacreBLEU and multi-eval. Our final en→de and de→en submissions achieve 40.48 and 48.75 SacreBLEU, 41.22 and 49.91 multi-eval respectively.

	en→de		de→en	
	SacreBLEU	multi-eval	SacreBLEU	multi-eval
last year's baseline	37.11	37.80	44.45	45.50
+ CL	40.11	40.89	47.77	48.89
+ DD, FT & ST	40.23	41.00	48.60	49.76
+ TEL	**40.48**	**41.22**	**48.75**	**49.91**

Table 2: BLEU scores of en→de and de→en NMT system on WMT22 biomedical test set.

6 Conclusion

This paper presents the submission of HW-TSC to the WMT23 biomedical translation task. We participate in en↔de language pair and perform a series of domain adaptation experiments based on the biomedical NMT system trained last year. The effectiveness of each domain adaptation method is demonstrated. Our experiments show that domain adaptation methods are effective for model training.

[2] https://github.com/moses-smt/mosesdecoder/tree/master/scripts/generic/mteval-v14.pl

References

Idris Abdulmumin. 2021. Enhanced back-translation for low resource neural machine translation using self-training. In *Information and Communication Technology and Applications: Third International Conference, ICTA 2020, Minna, Nigeria, November 24–27, 2020, Revised Selected Papers*, volume 1350, page 355. Springer Nature.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, page 489. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.

Ekaterina Garmash and Christof Monz. 2016. Ensemble learning for multi-source neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1409–1418.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Adam Lopez. 2008. Statistical machine translation. *ACM Computing Surveys (CSUR)*, 40(3):1–49.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30.

Xuan-Phi Nguyen, Shafiq Joty, Wu Kui, and Ai Ti Aw. 2020. Data diversification: a simple strategy for neural machine translation. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 10018–10029.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, page 186. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96. Association for Computational Linguistics (ACL).

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wei Wang, Ye Tian, Jiquan Ngiam, Yinfei Yang, Isaac Caswell, and Zarana Parekh. 2020a. Learning a multi-domain curriculum for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7711–7723.

Yiren Wang, Lijun Wu, Yingce Xia, Tao Qin, ChengXiang Zhai, and Tie-Yan Liu. 2020b. Transductive ensemble learning for neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6291–6298.

Daimeng Wei, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Xiaoyu Chen, Hengchao Shang, Jiaxin Guo, Minghan Wang, Lizhi Lei, Min Zhang, et al. 2021. Hw-tsc's participation in the wmt 2021 news translation shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 225–231.

Daimeng Wei, Zhiqiang Rao, Zhanglin Wu, Shaojun Li, Yuanchang Luo, Yuhao Xie, Xiaoyu Chen, Hengchao Shang, Zongyao Li, Zhengzhe Yu, et al. 2022. Hw-tsc's submissions to the wmt 2022 general machine translation shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 403–410.

Daimeng Wei, Zhanglin Wu, Hengchao Shang, Zongyao Li, Minghan Wang, Jiaxin Guo, Xiaoyu Chen, Zhengzhe Yu, and Hao Yang. 2023. Text style transfer back-translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 7944–7959, Toronto, Canada.

Zhanglin Wu, Zongyao Li, Daimeng Wei, Hengchao Shang, Jiaxin Guo, Xiaoyu Chen, Zhiqiang Rao, Zhengzhe Yu, Jinlong Yang, Shaojun Li, et al. 2023. Improving neural machine translation formality control with domain adaptation and reranking-based transductive learning. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 180–186.

Zhanglin Wu, Jinlong Yang, Zhiqiang Rao, Zhengzhe Yu, Daimeng Wei, Xiaoyu Chen, Zongyao Li, Hengchao Shang, Shaojun Li, Ming Zhu, et al. 2022. Hw-tsc translation systems for the wmt22 biomedical translation task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 936–942.

Hao Yang, Zhanglin Wu, Zhengzhe Yu, Xiaoyu Chen, Daimeng Wei, Zongyao Li, Hengchao Shang, Minghan Wang, Jiaxin Guo, Lizhi Lei, et al. 2021. Hw-tsc's submissions to the wmt21 biomedical translation task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 879–884.

Xuan Zhang, Pamela Shapiro, Gaurav Kumar, Paul McNamee, Marine Carpuat, and Kevin Duh. 2019. Curriculum learning for domain adaptation in neural machine translation. In *Proceedings of NAACL-HLT*, pages 1903–1915.

Investigating techniques for a deeper understanding of Neural Machine Translation (NMT) systems through data filtering and fine-tuning strategies

Lichao Zhu[3], Maria Zimina-Poirot[1]
Maud Bénard[1], Behnoosh Namdar[1], Nicolas Ballier[1 3], Guillaume Wisniewski[3], Jean-Baptiste Yunès[2]
[1]CLILLAC-ARP, [2]IRIF, [3]LLF
Université Paris Cité, F-75013 Paris, France
lichao.zhu, maria.zimina-poirot, maud.benard, behnoosh.namdarzadeh,
nicolas.ballier, guillaume.wisniewski, jean-baptiste.yunes@u-paris.fr

Abstract

In the context of this biomedical shared task, we have implemented data filters to enhance the selection of relevant training data for fine-tuning from the available training data sources. Specifically, we have employed textometric analysis to detect repetitive segments within the test set, which we have then used for refining the training data used to fine-tune the mBart-50 baseline model. Through this approach, we aim to achieve several objectives: developing a practical fine-tuning strategy for training biomedical in-domain fr<>en models, defining criteria for filtering in-domain training data, and comparing model predictions, fine-tuning data in accordance with the test set to gain a deeper insight into the functioning of Neural Machine Translation (NMT) systems.

1 Introduction

The objective of our contribution to the biomedical shared task this year is to gain deeper insights into the NMT training pipeline, assess the factors influencing performance, and evaluate the robustness of the training system.

Our training strategy was to build tailor-made fine-tuning data with regard to the test data. We calculated repeated segments (Salem, 1986) in the test data and used a selection of them to extract the corresponding data set from the available training data (as outlined in Section 2.1.1). In order to highlight the "reproducibility", we consistently adhered to the same pipeline with minimum settings presented in Section 2.

While our experience encountered various technical obstacles that likely affected our system's performance, these challenges prompted us to prioritize explainability and comparability. This became especially important when the system produced irrelevant results and hallucinations. We elaborated on this in Sections 3 and 4. The remaining sections of the paper are organized as follows: Section 2 outlines our filtering pipeline, Section 3 delves into our results, and Section 4 provides a discussion of these results. Section 5 explores related research and outlines future work.

2 Optimizing Fine-Tuning Data Selection and Pipeline

For this shared task, we exclusively fine-tuned the "mBart-50 large" baseline model (Tang et al., 2020) over 3-epoch, 5-epoch, and 10-epoch training cycles. Our approach prioritized reproducibility and involved a clear distinction between a "horizontal" dimension (facilitating inter-system analysis, especially in comparison with e-translation[1]) and a "vertical" dimension (examining differences in training strategies within the same system). The fine-tuned setting was established with the following parameters set as a minimum: "num_train_epochs", "train_file", "validation_file", "test_file", "per_device_train_batch_size"[2].

2.1 Raw Training Data

As recommended by the WMT biomedical shared task, we employed the corpora listed in Table 1. The entire corpus, intended for fine-tuning, comprises over 90M words in English and more than 100M in French. It is from these corpora that we extracted the filtered aligned sentences in both source and target languages.

2.1.1 Segmental Proximity Analysis for Training Data Filtering

Our training data filtering strategy is rooted in the theoretical principles of segmental proximity analysis. Repeated segments are sequences that are automatically identified as being repeated within

[1]https://commission.europa.eu/resources-partners/etranslation_fr

[2]Because of hardware constraints, we had to significantly reduce the batch dimension to 4

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 275–281
December 6–7, 2023.

Corpus	Lines	Words
PubMed abstract	13,033	2,429,484 (en) 3,051,103 (fr)
UFAL	2,693,509	89,191,554 (en) 100,024,568 (fr)
Edp	821	92,309 (en) 110,977 (fr)
Khresmoi	1,500	28,454 (en) 33,189 (fr)
Scielo[3]	9,393 9,501	213,684 (en) 262,377 (fr)

Table 1: Raw corpora used

UFAL (en)	30848
UFAL (fr)	16791
Pubmed abstract (fr)	8393
Pubmed abstract (en)	3566
Edp (fr)	54
Edp (en)	10
Khresmoi (fr)	19
Khresmoi (en)	1

Table 2: Matches in raw training corpus

the same text or across different texts.[4]

The computation of repeated segments (Salem, 1986) is a useful tool in corpus analysis. The computed inventory of repeated segments is of undeniable interest for discourse analysis (Sousa, 2014; Gledhill et al., 2017). This tool facilitates the examination of various discourse phenomena, encompassing the circulation of formulaic expressions, discourse routines, lexico-grammatical patterns, and more. On a cognitive level, the analysis of repeated segments can offer a substantial contribution to the study of knowledge pattern dissemination in specialized discourse. For these reasons, the computation of repeated segments holds substantial value in text profiling, especially in the context of training data selection. Our working hypothesis posits that discerning semantic proximity within a vast dataset can be accomplished through a deliberate selection of repeated segments guided by specific formal criteria, including factors such as frequency and segment length. The method relies on the assumption that related texts share common discourse properties, including phraseology, terminology, and structural patterns. These elements can be effectively "captured" through repeated segments computation and unveiled through segmental proximity analysis (Salem, 1986; Lebart et al., 1997).

The concept of segmental proximity has been thoroughly explored in the work of (Salem, 2006), where the statistical properties of this phenomenon were demonstrated. In this study, Salem (*ibid.*:1) *"considers measures of similarity based on the computation of the frequencies of identical sequences of words among the texts to be compared"*.

In order to identify common sequences within the two sections of a comparable monolingual corpus, it was first necessary to compile a list of segments containing a minimum of four words that were repeated in both parts of the corpus. All segments found exclusively in one part of the corpus were removed from consideration. The sequences that remained were then selected based on their length and their presence in each of the comparable parts of the corpus. By prioritizing relatively longer sequences, we were able to exclude many shorter and more frequent sequences, many of which consisted of common combinations of function words such as "of the" or "by means of."

According to the findings presented in Salem (Salem, 2006), the analysis of lexical distances and proximity indices computed on individual forms (such as the Jaccard index and Chi-square distance) did not reveal any significant affinity between the two sections of the comparable corpus under study. However, when calculations were based on the identification of repeated segments, it became evident that the two parts shared a relatively high number of extended sequences. This methodology enables the study of a range of phenomena related to the circulation of textual units that surpass individual vocabulary items.

Following this line of research, we compiled a systematic inventory of all repeated segments with a length of at least four words, such "acute respiratory syndrome coronavirus", "followed by maintenance therapy", etc. in each test set (English and French), which consisted of texts provided by WMT. We used *iTrameur* (https://itrameur.clillac-arp.univ-paris-diderot.fr) to facilitate this process. We then selected repeated segments with a total frequency of 10 or more. This curated inventory was used for the profiling and filtering of available medical text datasets (training data). We then selected repeated segments having a total frequency

[4]Consortium HN CORpus, Langues et Interactions - Huma-Num: https://corli.huma-num.fr/en/glossaire/repeated-segments/

of 10 or more and used this inventory for profiling and filtering of available medical text sets (training data). This process allowed us to build a dataset that shared common discourse properties and demonstrated semantic similarity with the test set.

2.2 Extraction of Aligned Sentences Containing Filtered Segments

By employing a list of repeated 4-word segments, we implemented a procedure to extract aligned sentences containing these filtered segments (Algorithm 1). To achieve this, we concatenated every delimiter (D) one by one (,;"~ &|#@='-.?!%*$()[]_:+«»§/) of *iTrameur*[5] and every word (g) of each 4-word repeated segment (G) of source language (S, which can be either English or French) to form a regular expression-like pattern GD. We used this pattern to match aligned sentences that contain the segment in both source (S_n, n is the index of matched sentence in raw training corpus of source language) and target (T_n) languages. The extraction result is reported in Table2. The extracted sentences are used to fine-tune the baseline model of mBart50.

Data: segment G, delimiter D and aligned sentences S and T
Result: aligned sentences S_n and T_n containing G

```
for g in G do
    GD ← concatenate(g,D)
    if GD in S_n then
        extract T_n
    end
end
```

Algorithm 1: Filtering and extraction algorithm

3 Results

In the absence of formal evaluation scores, our approach was largely based on textometric browsing techniques (Zimina, 2005) and qualitative analysis of our submitted translations. These translations were produced using a 3-epoch and 5-epoch training for the en-fr corpus, and a 3-epoch training for the fr-en corpus.

For example, Figure 1 shows a parallel section map generated by *iTrameur*, which helps visualize the alignment of parallel sections from two fr-en translations generated by mBart-50 baseline and mBart-50 5-epoch. The map highlights the presence or absence of the token "violence" in both translations. In the contexts displayed below the map, occurrences of repeated segments are underlined. We employ this tool to compare the translation outputs of various systems.

en-fr	Number	Frequency
train set (en)	60	482.144.115
test data (en)	41	98
fr-en	Number	Frequency
train set (fr)	70	176.229.164
test data (fr)	41	109

Table 3: The occurrence of 4-word sequences used for training in the train set corpus and test data (en, fr)

Texts in French	Number	Frequency
mBart50 (3-epoch)	2	9
mBart50 (5-epoch)	2	4
Texts in English	Number	Frequency
mBart50 (3-epoch)	0	0

Table 4: The presence of 4-word sequences (en, fr) in translations generated by our systems

In accordance with our training strategy, which is based on segmental proximity analysis (as described in Section 2.1.1), we expected the test data and the test set to have a substantial number of long sequences in common, assuming that many of the long sequences used for training would be shared. To confirm this, we examined the presence of the 136 four-word sequences (repeated segments) used in training within the test data. The results align with our research strategy, as shown in Table 3): 68% (41 sequences) of the English sequences are found in the en-fr test data and 59% (41 sequences) of the French sequences are present in the fr-en test data.

Continuing along this line of investigation, we examined our submitted translations, yet the analysis revealed that our translated texts had very little overlap with the repeated segments employed in training, as demonstrated in Table 4.

In the following paragraphs, we narrow our focus to two sequences taken from the repeated segments used for training. These examples help illustrate the challenges encountered by our systems and highlight the complexities involved in drawing

[5] https://itrameur.clillac-arp.univ-paris-diderot.fr

Figure 1: Parallel section map generated by *iTrameur*.

significant conclusions from a qualitative analysis of our translations.

With 13 occurrences in the fr-en test data, the sequence *"violence envers les femmes"* presents an intriguing case of terminology. A closer examination of the raw training corpus reveals that several other expressions are commonly employed in French to express the same concept, such as *"violence faite aux femmes"* and *"violence contre les femmes"*, among others. In English, we note a reduced degree of variation, primarily using "violence against women" and, to a lesser extent, "violence directed against women", which is not as prevalent as the former. According to the European Institute for Gender Equality (EIGE), an EU agency, the most accurate translation is "violence against women".[6]. However, in the translations generated by the 3-epoch training, this accurate translation is absent, and the proposed translations ("the women's domestic violence" and "gender-based violence") do not correspond to the same concepts. The medical term "blood flow" appears 10 times in the en-fr test data and is part of one of the repeated segments used for training, specifically "blood flow in dogs." It is also a frequent component in complex noun phrases, including "the dermal blood flow," "regional blood flow," or "auricular dermal blood flow" (with 8 occurrences). Both the 3-epoch and 5-epoch systems encounter challenges when translating this repeated segment and the complex noun phrases. Firstly, there's a high degree of terminological variation in the French texts (*"écoulement sanguin"*, *"débit sanguin"*, *"flux sanguin"*, *"circulation sanguine"*), given the limited occurrences of the term in English. Additionally, we observe instances of hallucinations and incomplete outputs in the 3-epoch system. The 5-epoch system, on the other hand, omits one element in the translation of the complex noun phrase, even though the 3-epoch system accurately translated it.

In a specific case, "regional blood flow," the 5-epoch system incorrectly deduced the semantic relationship between the head and a modifier, yielding "*l'écoulement régional du sang*" instead of the correct "*l'écoulement sanguin régional*", which the 3-epoch system produced.

4 Discussion

4.1 Errors and Inconsistencies Arising From Variations in the Train Set

In general, the output exhibits numerous errors and inconsistencies primarily arising from terminological variations within the train set and the inherent

[6] https://eige.europa.eu/publications-resources/thesaurus/terms

heterogeneity of the selected training data. For instance, the terms "gender-based violence" and "violence against women" are both employed in comparable contexts within the train set, as illustrated by segments like "Many women who experience gender-based violence may never seek any formal help..." and "Violence against women is a global phenomenon" (source: PubMed abstracts, train set: en).

4.2 Hallucinations with mBart-50

Analyzing the translations generated by mBart-50 at different epochs (5 and 10) proves to be an interesting area of research. According to Lee et al. (2018), hallucinations occur when the model produces significantly different and inadequate outputs when the source is subjected to specific noise models. Therefore, we can suggest that there may be instances where the model ceases to translate the source text and instead generates an output composed solely of a continuous sequence of tokens from the present invention. This could be seen as an alternative form of hallucination in epoch 10. Moreover, in epoch 10, there are also examples of incomplete translations produced by the system. Based on our analysis, we observed that these errors tend to be resolved during the training process. Consequently, in mBart-50 (5-epoch) fr-en, there are no "X" tokens present, in contrast to the baseline model translation. This result is highlighted by the calculation of generalized co-occurrence networks conducted using *iTrameur*. Figures 2 and 3 depict co-occurrence networks that represent the most characteristic lexical attractions in the fr-en translations produced by two models: mBart-50 baseline and mBart-50 5-epoch. The numbers on the edges represent the strength of lexical attraction: *Specificity Index* > *9* (Co-frequency > 1).[7] Co-occurrence networks serve as a monitoring tool for tracking changes across various training stages.

5 Related Research and Future Work

Corpus filtering is discussed in many previous and recent works for training data preparation with different approaches : perplexity threshold of text segments(Moore and Lewis, 2010), metric evaluations of raw NMT models' outputs (Duh et al., 2013), acceptability of filtering evaluated by mulitlingual BERT classifier(Zhang et al., 2020), etc. Our approach aims at data relevance between a given test set and in-domain training data.

5.1 Robustness of NMT Models

An essential aspect to consider is that the system generates tokens regardless of whether it possesses the relevant information for translation. However, this raises the need for potential trigger warnings in situations where the system lacks adequate data for accurate translation. This suggests an avenue for developing a confidence index that reflects the system's efforts when generating output. We consider to explore various parameters based on sentence level and on a token level to build such a confidence index, e.g. the scores used by certain large language models to assess the confidence of each token's projection or beam search and the number of competitors for each token to gauge the complexity of a text for translation.
Another aspect to consider is the model's over/underfitting. By plotting our training and test data features in Figure 4, we find out that the model is indeed overfitted from depth 4. That explains partially why the model under-performed and helps us to choose a better data fitting strategy in the future.

5.2 Fine-tuning of mBart-50 and Other Multilingual Systems

For the moment, we only tried to fine-tune mBart-50 as a multilingual large language model, whereas other systems have been developed since, some of them with many more parameters. We may try to replicate or fine-tune experiments with more classical systems such as SYSTRAN Model Studio Advanced (https://u-paris.fr/plateforme-paptan), but also other different multi-lingual large language models such as mT5 (Xue et al., 2020) or Bloom (Scao et al., 2022), a 176-billion parameter language model, in spite of its carbon footprint (at least 24.7 tons of carbon just for the dynamic power consumption) (Luccioni et al., 2022).

Finally, it is worth noting that our fine-tuning efforts were primarily centered around mBart-50, a multilingual large language model. However, since our experiments, various other systems have emerged, some with significantly more parameters. It might be advantageous for us to replicate or fine-tune experiments with more conventional systems based on translation models, such as the SYSTRAN Model Studio Advanced (available at Université Paris Cité: https://u-paris.fr/plateforme-paptan), and explore different multilingual large

[7] For specific details regarding the computation of *Specificity Index*, refer to (Lebart et al., 1997).

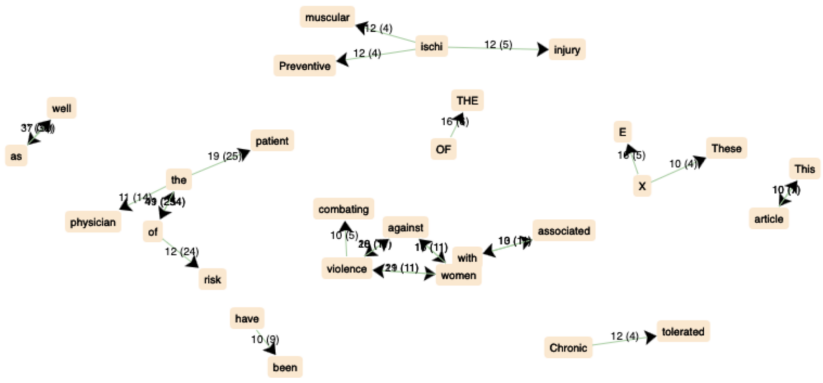


Figure 2: Co-occurrence networks for the baseline.

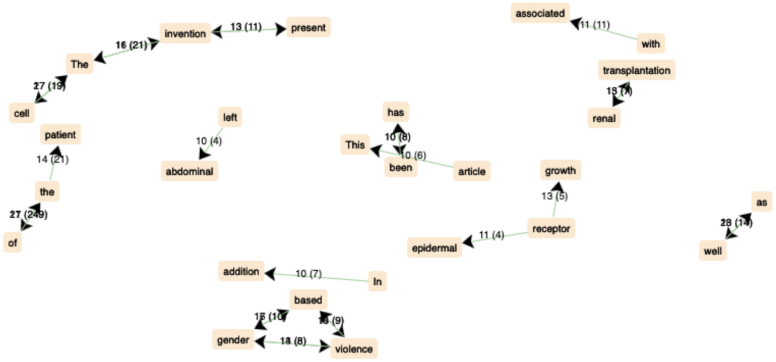


Figure 3: Co-occurrence networks for the 5-epoch system.

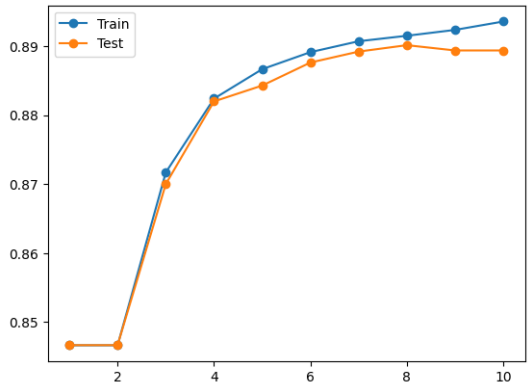


Figure 4: Model's accuracy prediction by training and test data features classification

language models like mT5 (Xue et al., 2020) or Bloom (Scao et al., 2022). Notably, Bloom is a huge 176-billion-parameter language model of major interest, despite its substantial carbon footprint, which amounts to at least 24.7 tons of carbon emissions solely for dynamic power consumption (Luccioni et al., 2022).

6 Conclusion

In this paper, we have described the translation systems used for the submissions in the WMT23 biomedical task6 (our data are available at: https://github.com/lichaozhu/WMT23). Nevertheless, due to certain hardware constraints, we were unable to pinpoint the exact reasons for the model's underperformance.

We also considered our previous participation in the biomedical task. Since 2021, we have recognized that having scores provided in advance and reference texts used for score computation can significantly facilitate our work. These resources enable a more critical evaluation of the translations we generate.

To address the absence of reference translations and evaluation results, translations can undergo

spot checks. In our work, these checks involved the use of qualitative examples to assess the model's successes and failures. Additionally, textometric browsing helped to unveil distinctive features within multiple machine translation outputs.

Acknowledgements

This publication has emanated from research supported in part by a 2021 research equipment grant (PAPTAN project)[8] from the Scientific Platforms and Equipment Committee, under the ANR grant (ANR-18-IDEX-0001, Financement IdEx Université de Paris).

References

Kevin Duh, Graham Neubig, Katsuhito Sudoh, and Hajime Tsukada. 2013. Adaptation data selection using neural language models: Experiments in machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 678–683, Sofia, Bulgaria. Association for Computational Linguistics.

Christopher Gledhill, Maria Zimina-Poirot, and Stéphane Patin. 2017. Lexico-grammaire et textométrie : identification et visualisation de schémas lexico-grammaticaux caractéristiques dans deux corpus juridiques comparables en français. *Corpus*.

Ludovic Lebart, André Salem, and Lisette Berry. 1997. *Exploring Textual Data*, volume 4. Springer Science & Business Media.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in neural machine translation.

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2022. Estimating the carbon footprint of bloom, a 176b parameter language model. *arXiv preprint arXiv:2211.02001*.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden. Association for Computational Linguistics.

André Salem. 2006. Proximités segmentales. In *Actes des 8e Journées internationales d'Analyse statistique des Données Textuelles (JADT 2006)*, pages 839–849, Besançon, France. Université de Franche-Comté.

André Salem. 1986. Segments répétés et analyse statistique des données textuelles. *Histoire & Mesure*, 1(2):5–28.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Serge de Sousa. 2014. De la statistique textuelle à l'analyse des idéologies: l'exemple du discours révolutionnaire en amérique latine (1810-2010). *Corela. Cognition, représentation, langage*, HS(15).

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Boliang Zhang, Ajay Nagesh, and Kevin Knight. 2020. Parallel corpus filtering via pre-trained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8545–8554, Online. Association for Computational Linguistics.

Maria Zimina. 2005. Bi-text Topography and Quantitative Approaches of Parallel Text Processing. In *The Corpus Linguistics 2005 conference*, volume 1 Issue 1, Birmingham, United Kingdom. Centre for Corpus Research, Birmingham University.

[8] Plateforme pour l'apprentissage profond pour la traduction automatique neuronale, in English: Deep Learning for Machine Translation at Universite Paris Cité. See the description of the platform on the project website: https://u-paris.fr/plateforme-paptan

MAX-ISI System at WMT23 Discourse-Level Literary Translation Task

Li An* **Linghao Jin*** **Xuezhe Ma**
University of Southern California, Information Sciences Institute
{lan72605, linghaoj}@usc.edu
xuezhema@isi.edu

Abstract

This paper describes MaxLab - Information Sciences Institute (MAX-ISI) Translation systems for the WMT23 shared task. We participated in the discourse-level literary translation task - constrained track. In our methodology, we conduct a comparative analysis between the conventional Transformer model and the recently introduced MEGA model, which exhibits enhanced capabilities in modeling long-range sequences compared to the traditional Transformers. To explore whether language models can more effectively harness document-level context using paragraph-level data, we took the approach of aggregating sentences into paragraphs from the original literary dataset provided by the organizers. This paragraph-level data was utilized in both the Transformer and MEGA models. To ensure a fair comparison across all systems, we employed a sentence-alignment strategy to reverse our translation results from the paragraph-level back to the sentence-level alignment. Finally, our evaluation process encompasses sentence-level metrics such as BLEU, as well as two document-level metrics: d-BLEU and BlonDe.

1 Introduction

This paper introduces our submissions to the WMT23 Shared Task: Discourse-Level Literary Translation (Zh-En), Constrained Track. Our submission comprises three translation systems: a primary system employing a paragraph-level transformer, a first contrastive system utilizing a sentence-level transformer, and a paragraph-level Mega model as the second contrastive system.

Until very recently, the predominant focus of context-aware Neural Machine Translation (NMT) research has been on parallel datasets that align at the sentence level, such as IWSLT17 (Cettolo et al., 2017) and OPUS (Tiedemann, 2012). More recent research endeavors have concentrated on literary translation, which is typically more intricate and requires the models to be able to capture long-range context for high-quality translations. For example, Thai et al. (2022) introduced the first multilingual paragraph-aligned dataset PAR3, sourced from public-domain non-English literary works.

We use Transformer as the baseline model. In order to assess whether a more advanced model can excel in modeling long-range sequences using literary data, which contains richer contextual information, we also include the MEGA (Ma et al., 2023) model for comparison. The foundational model architectures we employ are introduced in Section 2.

In Section 3, we provide an extensive explanation of our systems. Within this section, Section 3.1 outlines the data pre-processing step. In this phase, we construct both sentence-level data, which comprises the filtered original data, as well as paragraph-level data. It's worth noting that aligning sentences in literary translation is not always feasible due to the possibility of sentence merging or truncation during the translation process. At the paragraph level, language models can adeptly exploit document-level context, resulting in a reduction of translation errors at the discourse level, as corroborated by human evaluations (Karpinska and Iyyer, 2023). Building on these encouraging findings, we created a dataset aligned at the paragraph level by aggregating multiple sentences from the provided literary dataset. Then, we propose three systems and evaluate those systems with both sentence-level and document-level metrics.

Section 4 presents the results that culminate in our final submissions. Additionally, we discussed challenges we encountered regarding discourse-level translation in Section 5.

*Equal contribution.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 282–286
December 6–7, 2023.

2 Model Architectures

We select the following two model architectures for our systems, taking into account their strong performance in the context of context-aware machine translation.

Transformer The Transformer architecture, as introduced by Vaswani et al. (2017), utilizes an encoder-decoder framework, leveraging a self-attention mechanism. This mechanism enables each position within a given sequence to interact with every other position, facilitating the computation of a comprehensive representation for the entire sequence.

In all our experiments, we employ the Transformer `base` version which consists of 6 encoder layers, 6 decoder layers, a model dimension of 512, and a FFN hidden dimension of 2048.

MEGA The recently unveiled MEGA (Moving Average Equipped Gated Attention) (Ma et al., 2023), addresses two longstanding limitations of the conventional Transformer model, which have impeded its performance on tasks involving long sequences. These limitations pertain to a weak inductive bias and a quadratic computational complexity.

MEGA employs a multi-dimensional, damped exponential moving average (Hunter, 1986) (EMA) in conjunction with a single-head gated attention mechanism to preserve inductive biases. Importantly, MEGA can replace the attention mechanism within the Transformer framework. Additionally, MEGA is of comparable size to the Transformer.

In total, the Transformer architecture is around 75M parameters; the MEGA architecture is around 77M parameters.

3 System Overview

3.1 Data Preprocessing

We first perform the following filtering steps on the training data:

- Remove translators' notes.
- Merge dialogues with tags "#<#" and "#>#" into one instance.
- Combine blank lines with their following line.

We construct sent-level and paragraph-level datasets separately.

Sentence-level dataset is constructed using the sentence alignment information, which is thoughtfully provided.

Paragraph-level dataset Considering the critical role played by context, particularly in literary translation, we further construct a paragraph-aligned corpus. This corpus is established based on the sentence alignment, allowing us to leverage context more effectively in our translations.

Data for each language pair is then encoded and vectorized with byte-pair encoding (Sennrich et al., 2016) using the `SentencePiece` (Kudo and Richardson, 2018) framework. We use separate vocabularies of size 32K for each language Zh and En.

Full corpus statistics are in Table 1.

Subset	Sent-level	Paragraph-level
Train	1742150	290315
Valid1	711	154
Valid2	810	148

Table 1: Instance counts across train and valid subsets.

3.2 System Architectures

Transformer-256 Our primary system employs a `Transformer-base` model at the paragraph level. Prior to tokenization, we structured the data into paragraphs, each with a maximum length of 256 characters on the source side (Zh). The model is subsequently trained and utilized for decoding on the paragraph-aligned corpus mentioned above.

Transformer-Sent In contrast, we conduct training for the transformer-base model using the sentence-level corpus.

MEGA-256 We adopted our proposed paragraph-aligned data as it demonstrates competitive efficacy in comparison to conventional Transformers across established benchmarks, including the LRA dataset, all while maintaining a significantly leaner parameter configuration.

3.3 Training

We train all models on the `fairseq` framework (Ott et al., 2019). All models were trained on 4 NVIDIA A40 GPUs. Following Vaswani et al. (2017); Fernandes et al. (2021), we use the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.98$, a linear decay learning rate scheduler with an initial value of 10^{-4},

System	Subset	BLEU	d-BLEU	BlonDe				
				all	pron.	entity	tense	d.m.
Transformer-Sent	VALID1	**26.40**	**26.40**	37.87	74.40	36.95	69.98	67.22
	VALID2	**16.40**	16.10	29.89	67.34	**49.05**	70.76	52.78
Transformer-256	VALID1	21.90	26.20	**40.92**	**84.17**	**40.47**	**78.72**	**72.71**
	VALID2	13.60	**16.30**	**33.50**	**79.72**	46.44	**81.57**	**68.80**
MEGA-Sent	VALID1	25.00	25.00	37.03	73.55	36.32	68.81	66.55
	VALID2	16.20	15.80	29.54	67.18	47.30	69.95	54.21
MEGA-256	VALID1	22.40	23.90	39.74	81.14	**40.47**	77.29	71.47
	VALID2	13.20	15.80	32.90	77.37	48.17	81.13	66.80

Table 2: Automatic metric results on the valid1 and valid2 sets. All reported BlonDe scores are F1s; pron. stands for pronoun, d.m. stands for discourse marker.

System	Sent-Level				Doc-Level	Human Annotator
	BLEU	chrF	COMET	TER	d-BLEU	Average
Transformer-256	34.1	53.3	78.24	**62.4**	**45.1**	73.59
Transformer-Sent	**34.5**	**54.7**	**79.14**	62.7	44.9	✗
MEGA-256	33.1	52.4	77.84	63.6	44.4	✗

Table 3: Automatic metric results of our submissions on the test set and the average score by different annotators on one sampled document. (Wang et al., 2023).

and increasing to $5e^{-4}$ during a warm-up phase of 4000, and a dropout of 0.2. We run inference on the validation set and save the checkpoint with the best BLEU score.

3.4 Post-processing

Since the final submission requires that each line must be aligned with the corresponding input line in the output files, we add this post-processing step to reverse our paragraph-level translation result to sentence-level alignment. We will discuss this further in the conclusion part.

Sentence-Alignment

1. Use the translated results at the sentence level as a reference

2. Calculated the similarity between each sentence in the translated paragraph and the *M* nearest sentences in the sentence-level translation

3. Align each sentence to the most similar one using Jaccard similarity on N-gram overlap as the similarity metric

3.5 Evaluation

To evaluate the discourse-level translation ability of three systems, we compute three metrics:

BLEU (Papineni et al., 2002) sentence-level BLEU is the most commonly used metric to evaluate the quality of machine-generated translations. We report the standard BLEU score calculated using `sacreBLEU` (Post, 2018)[1] in our systems.

d-BLEU (Liu et al., 2020) document-level sacreBLEU is computed by matching n-grams in the whole document. Note that all evaluations are case-sensitive

BlonDe (Jiang et al., 2022) is introduced as a document-level automatic metric that calculates the similarity-based F1 measure of discourse-related spans across four categories (*pronoun, entity, tense and discourse marker*).

4 Results

The results of our experiments are presented in Table 2. We evaluate our models on the provided two validation sets and list model performances

[1]The sacreBLEU signature is BLEU+case.mixed+lang.src-tgt+numrefs.1+smooth.exp+{test-set}+tok.13a.

on three automatic metrics, i.e., BLEU, d-BLEU, and BlonDe. Given that BLEU scores compare n-grams on a sentence-level basis, we extend our evaluation to encompass d-BLEU and BlonDe metrics, providing a comprehensive assessment of the models' proficiency in discourse-level translation. The results of the test set are presented in Table 3.

Transformer *vs.* MEGA As per the outcomes presented in Table 2, Transformer models slightly surpass MEGA models in both sentence-level and paragraph-level translations. While MEGA demonstrates superior capabilities in long-range sequence modeling, its limited enhancement may be attributed to the fact that current data are not lengthy and doesn't capture sufficient useful context (Jin et al., 2023). Furthermore, the discrepancy in BLEU scores is more pronounced than the variation in BlonDe scores.

Sent-level *vs.* Paragraph-level Based on the results presented in Table 2, there is a discrepancy between BLEU and BlonDe evaluations. Specifically, it is observed that sentence-level translation exhibits a better performance in terms of the BLEU metric, whereas paragraph-level models demonstrate a substantial improvement when assessed using the BlonDe metric.

As delving into the four distinct categories in BlonDe, a consistent trend of enhancement emerges across each category with the adoption of paragraph-level translation. Particularly, marked improvements are observed within the pronoun and tense categories. This can be attributed to the inherent reliance of pronouns and tenses on contextual information. These empirical results demonstrate that paragraph-level data provides more useful contextual signals than sentence-level data.

5 Discussion

Limitation of sentence alignment Literary texts often rely on context that spans beyond individual sentences, making strict sentence alignment impractical. As evidenced in our results, paragraph-level translation excels in preserving contextual information, like pronouns and tenses. However, the insistence on maintaining sentence-level alignment imposes constraints on model selection, hindering flexibility and adaptability.

Limitation of evaluation metrics The current evaluation metrics are not capable enough of measuring document-level machine translation. The most commonly used metric, BLEU, and its variant, d-BLEU, may struggle to fully capture the context awareness and coherence that is crucial at the document level translation.

6 Conclusion

This paper describes the submission to the WMT23 literary translation shared task - constrained track. We compare traditional Transformer models to the newer MEGA model, integrating paragraph-level data into both. Transformer models outperform MEGA in both sentence and paragraph translation on this literary dataset. We observe a discrepancy between BLEU and BlonDe evaluations, with the latter favoring paragraph-level translation. These results emphasize the challenges of document-level translation and the importance of more context-aware evaluation metrics.

References

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Niehues Jan, Stüker Sebastian, Sudoh Katsuitho, Yoshino Koichiro, and Federmann Christian. 2017. Overview of the iwslt 2017 evaluation campaign. In *Proceedings of the 14th International Workshop on Spoken Language Translation*, pages 2–14.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. 2021. Measuring and increasing context usage in context-aware machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6467–6478, Online. Association for Computational Linguistics.

J. Stuart Hunter. 1986. The exponentially weighted moving average. In *Journal of Quality Technology*.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States. Association for Computational Linguistics.

Linghao Jin, Jacqueline He, Jonathan May, and Xuezhe Ma. 2023. Challenges in context-aware neural machine translation.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation.

Xuezhe Ma, Chunting Zhou, Xiang Kong, Junxian He, Liangke Gui, Graham Neubig, Jonathan May, and Luke Zettlemoyer. 2023. Mega: Moving average equipped gated attention. In *The Eleventh International Conference on Learning Representations*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, Weiyu Chen, Yvette Graham, Bonnie Webber, Philipp Koehn, Andy Way, Yulin Yuan, and Shuming Shi. 2023. Findings of the WMT 2023 shared task on discourse-level literary translation. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*.

The MAKE-NMTViz System Description for the WMT23 Literary Task

Fabien Lopez[1] and **Gabriela Gonzalez-Saez**[1] and **Damien Hansen**[1 6] and **Mariam Nakhle**[1 5] and **Behnoosh Namdarzadeh**[3] and **Marco Dinarelli**[1] and **Emmanuelle Esperança-Rodier**[1] and **Sui He**[4] and **Sadaf Mohseni**[3] and **Caroline Rossi**[2] and **Didier Schwab**[1] and **Jun Yang**[4] and **Jean-Baptiste Yunès**[3] and **Lichao Zhu**[3] and **Nicolas Ballier**[3]

1 Univ. Grenoble Alpes,
CNRS, Grenoble INP, LIG
38000 Grenoble, France
2 Université Grenoble Alpes
3 Université Paris Cité
4 Swansea University
5 Lingua Custodia, France
6 Université de Liège, CIRTI, 4020 Liège, Belgique
Contact: fabien.lopez@univ-grenoble-alpes.fr

Abstract

This paper describes the MAKE-NMT-Viz's submission to the WMT 2023 Literary task. As a primary submission, we fine-tune the mBART50 model using Train, Valid1, and Test1 as part of the GuoFeng corpus (Wang et al., 2023b).We followed similar training parameters to (Lee et al., 2022) when fine-tuning mBART50. For our contrastive1 submission, we used a context-aware NMT system based on the concatenation method (Lupo et al., 2022). The training was performed in two steps: (i) a traditional sentence-level transformer (Vaswani et al., 2017) was trained for 10 epochs using GeneralData, Test2, and Valid2; (ii) second, we fine-tuned such Transformer using document-level data, with 3-sentence concatenation as context, for 4 epochs using Train, Test1, and Valid1 data. We then compared the three translation outputs from an interdisciplinary perspective, investigating some of the effects of sentence- vs. document-based training. Computer scientists, translators and corpus linguists discussed the remaining linguistic issues for this discourse-level literary translation.

1 Introduction

In order to analyse literary translations, we have gathered an interdisciplinary team of translators, linguists and computational scientists. We used this opportunity to explore neural machine translation of literary texts as a test set for test suites and unsolved issues for MMT literary translations, especially for the Chinese-English language pair. While the topic of literary machine translation has gained momentum in the last years, there have still been few attempts to customize systems to literary data, although this idea is also drawing attention (Kenny and Winters, forthcoming). Indeed, research has been carried out on this subject, notably on Catalan (Toral and Way, 2018), but also on Slovenian (Kuzman et al., 2019), German and Russian (Matusov, 2019), and on French (Besacier and Schwartz, 2015), where research suggests that MT systems can be further fine-tuned on specific genres and individual translator styles (Hansen and Esperança-Rodier, V.2023).

Of course, these very attempts bring about many issues concerning textual ownership, copyright, translator status and livelihood, possibly lowered quality, cognitive friction, etc. (Taivalkoski-Shilov, 2019). It is therefore important to include these ethical aspects into the research and clarify its objectives: for instance, whether MT should serve as a reading aid (Oliver González, 2017), or as a post-editing tool that may decrease the effort needed to translate (Kolb, 2020) and constrain creativity (Guerberof-Arenas and Toral, 2022).

Part research has also focused on evaluating the use of existing tools for literary texts. In the context of Chinese to English, attention has been paid to some of the specific shortcomings of MT systems, such as the translation of adjectival possessive pronouns (Jiang and Yu, 2017), or theme-rheme progressions (Jiang and Niu, 2022). Such limitations can indeed have a drastic impact on readers' acceptance, which Shih (2016) explores in the context of online folktales, confirming that the text's function plays a large role in this respect.

Lastly, Thai et al. (2022) have also pointed the incompatibility of MT metrics, document-level or otherwise, for literary texts, concluding that "hu-

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 287–295
December 6–7, 2023.

man expert evaluation is currently the only way to judge the quality of literary MT".

The rest of the paper is organised as follows: Section 2 details our approaches to the task and the training data of our experiments, Section 3 presents the results and Section 4 discusses them.

2 Data and Tools Used

This section details the toolkits we used and our training data for the three submissions authorised for the task. We first used part of the training data proposed by the organisers (Wang et al., 2023a) to observe the translations from mBART50 from Chinese into English before fine-tuning mBART (primary submission). We then used a fine-tuned context-aware concatenation-based Transformer trained at document level (contrastive1 submission) and a traditional sentence-level Transformer (contrastive2 submission).

2.1 Primary model: mBART50 fine-tuning

As a primary submission, we used GuoFeng corpus (Wang et al., 2023a) to fine-tune the mBART50 model with Chinese-English data, using the Train set for training, Test1 as test set, and Valid1 as validation set. We followed similar training parameters to (Lee et al., 2022) when fine-tuning mBART50. As (Lee et al., 2022), we trained for 3 epochs, using gelu as an activation function, with a learning rate of 0.05, dropout of 0.1 and a batch size of 16 (we parallelised two A100 GPUs with batch size 8 per device). We decoded using a beam search of size 5.

2.2 Contrastive models

We submit two contrastive models, the first is a context-aware model (*contrastive1*) built on the second system, a sentence-level model (*contrastive2*).

For our contrastive1 submission, we used a context-aware NMT system based on the concatenation method (Lupo et al., 2023). The training was performed in two steps: (i) a sentence-level transformer (Vaswani et al., 2017) was trained for 10 epochs[1] using General Data as train set, Test2 as test set, and Valid2 as validation set ; (ii) second, we fine-tuned at document-level using 3-sentence concatenation for 4 epochs[2] using Train as train set, Valid1 as validation set and Test1 as test set. During the fine-tuning, we used ReLU as an activation function, with an inverse square root learning rate decay, dropout of 0.1, and a batch size of 64. We decoded using a beam search of size 4. For our contrastive2, we used the model trained at step (i) (sentence-level). The training parameters were an inverse square root learning rate decay, a dropout of 0.1, and a batch size of 64. We decoded using a beam search of size 4.

[1]We used only 10 epochs because of time constraints

[2]We used only 4 epochs because of time constraints

2.3 Evaluation Metrics

To evaluate our models, we use the BLEU score metric (Papineni et al., 2002) as implemented in the Moses package.

We performed a human annotation of errors in the translation obtained by our primary submission. 109 segments were selected and annotated by three evaluators that are Chinese native speakers. To measure the inter-annotator agreement, we used Fleiss' kappa (Fleiss et al., 1971). The score is calculated to measure the inter-rater reliability of the annotations as the following equation

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where $P_o - P_e$ measures the real concordance of annotations that are not achieved above chance, while $1 - P_e$ measures the achievable concordance of annotations above chance. In our case, we computed errors by type as well as error types by segment (6 types and 109 segments, cf4.3.2).

3 Experiments and Results

We provide a human analysis of the primary model by discussing the improvements observed with the mBART fine-tuning with respect to the baseline. Additionally, we report the BLEU scores of our three systems.

3.1 Baseline of primary model: mBART50

During the training phase of the competition, with the standard HuggingFace implementation of mBART50, we observed the following issues when we translated Test1 from Chinese to English, which was part of the data provided for training by the organisers:

- hallucinations
- discrepancy between the Chinese input and the English translations
- tense concord
- co-referentiality issues for pronouns

Most textual discrepancies between the sizes of the sentences in the two languages were fixed by the fine-tuning as well as hallucinations and Chinese characters in the English translations. We nevertheless noticed a certain number of Chinese characters in the mBART50 translations, which decreased after our fine-tuning, and we only found 18 examples for all the 16,742 sentences, mostly for the fantasy genre, when referring to named entities or specific attributes of the universe (*Skills: Blade Technique, Wing Protection,*).

3.2 Fine-tuning with Literary Data

In this section, we analyse the outputs qualitatively. This analysis consists of an initial description of the baseline and fine-tuned outputs, followed by a deeper examination of the syntactic and semantic functions of the produced outputs by both models.

Instances of hallucinations were observed in the outputs of our baseline model. The hallucinated elements are present in the source text, so they are not elements which are not present in the source text. According to Lee et al. (2018), hallucinations can be defined as the model producing a vastly different and inadequate output when the source is perturbed under a specific noise model. Thus, we may suggest that there exist other instances where the model ceases translation of the source text and proceeds with generating output punctuated solely by a sequence of continuous commas (,,,,,,,,,), which may represent an alternative manifestation of hallucination. Interestingly, it is noteworthy that the fine-tuned outputs did not exhibit any instances of hallucination. However, it should be mentioned that few Chinese tokens were observed in the fine-tuned outputs. In the Chinese source text, the equivalent of the word "businessmen" is placed at the left periphery of the sentence, having a pragmatic effect that involves topic introduction or re-introduction, based on the context. Both the baseline and fine-tuned models take the left dislocated element to the right periphery of the sentence, thereby inducing an alternation in the sentence's intended meaning. As we observed, the baseline models chunk the sentences and use commas instead of employing coordinations, relative clauses, or more complex structures. In this example, the baseline model produces "Ten minutes later. consciousness is exhausted. scattered" by separating each chunk or even token with a period. In contrast, the fine-tuned model generates "Ten minutes later, his consciousness was exhausted and dissipated.", using coordination to form a united sentence. This represents another instance of the fine-tuned model's proficient manipulation of structures, wherein it employs a relative clause "which" to interconnect the sentences. Ex: "Wang lived in the 413 bedrooms of the West school district, Lins lived in the 413 bedrooms of the East school district." Fine-tuned: 09primary: "Wang Yicheng stayed in 413, which was in the West campus. Lin Sisi stayed in 413, which was in the East campus." Furthermore, the choice of tense seems to be different in the two models: As for the fine-tuned model, a preference for the past tense becomes evident. Conversely, as for the baseline model, an over-use of the present tense is observed in its outputs. We may also add that baseline models tend to favour the indicative mood, which indicates assertion, as seen in an example like "What's wrong with the game?". On the other hand, fine-tuned models have been *trained* to produce sentences in moods that exhibit a reduced level of assertiveness, as evidenced by constructions like "Could there be a problem with the game?".

3.3 BLEU scores

In this section, we report the results of our primary, constrastive1, and contrastive2 in terms of BLEU score computed using Test1 and Test2 datasets at the end of the full training process of each model. The official results of the competition on test3 were not computed as the reference translations were not provided (at the time of writing this article).

Model	Test1	Test2
primary	22.31	–
contrastive1 (document-level)	19.03	17.58
contrastive2 (sentence-level)	22.31	18.22

Table 1: BLEU score for primary, contrastive1 and contrastive2 systems.

Table 1 shows that our primary system achieves the same BLEU score as contrastive2[3], the sentence-level transformer implementation. We notice that the document-level system (contrastive1) is not better than the sentence-level model. This

[3]Primary and contrastive2 scores on Test1 are identical due to coincidence.

might be explained by the few epochs used for training.

4 Discussion

4.1 Lexical Complexity

To appreciate the relative complexity of the terms used in the translations we first qualitatively compared the translations and contrastive2 seemed to be more elaborate, so we tested this impression with more quantitative means. We investigated the vocabulary growth curves of the three translations using the functions available from the languageR package (Baayen and Shafaei-Bajestan, 2019) to find out that the number of different types progress on the same rhythm for the different translations. In this type of representation, the horizontal axis corresponds to the expansion of the translation corpus (number of tokens) and the vertical axis corresponds to the number of types. The first lower series of curves corresponds to the number of hapaxes. As can be seen in Figure 1, the progression is very similar for the different translations we produced, while the mBART fine-tuning translation (primary) seems to be more verbose as the translation contains pore tokens than the two contrastive translations. The difference between our different models is clearly not lexical.

4.2 Challenging Literary Aspects of the Test Set

The first challenge was the size of the testing data, which resorted to different text genres, but was 30 times bigger than other challenge datasets like for the biomedical task in 2021. An additional difficulty was the paucity of metadata for the 14 genres or for chapter attributions (22 announced and 12 found).

4.3 Translation Quality analysis based on Error Annotation

4.3.1 Quality overview

In total, 109 sample segments were randomly selected from the twelve translated texts generated by the fine-tuned mBART50 model. Based on these sample segments, each translated text was assigned an overall grade individually by three annotators on a scale of 1 to 10, with 1-3 denoting "Very Poor", 4-6 denoting "Poor", 7-8 denoting "Moderate", and 9-10 denoting "Good". The annotators are native Chinese speakers with near native level of English competence. They work in the domain of translation training and linguistics with an advanced proficiency of Chinese-English translation. The three grades given by the annotators for each text were then averaged to obtain a relative ranking of each translation. Overall, the twelve translations achieved an average score of 5 out of 10 in general, with a standard deviation of 0.87. Specifically, seven subgenres were identified among the twelve texts, namely: fantasy (4 texts), ancient romance (2 texts), military (1 text), thriller (1 text), modern romance (2 texts), sci-fi (1 text), and online games (1 text). All the sub-genres are typical in contemporary web novels. Notably, there is not a clear cut between different sub-genres and this categorisation is for analytical purposes only. Among the identified subgenres, the ranking from high to low quality is as follows: thriller (6.0 out of 10), fantasy (5.7), online games (5.4), sci-fi (5.0), ancient romance (4.7), modern romance (4.6), and military (3.8). While subgenre types might be a factor in influencing the quality of the translation given their language styles (e.g., the proportion of conversational segments, terminologies, formality, etc.), this line of discussion requires further evidence. Among the sample segments, the quality and language style of individual source text seem to play a more vital role in the overall quality of the translations. Several prominent error types linking to the stylistic features of the texts were identified, as detailed below.

4.3.2 Error typology

To obtain a more detailed insight into the quality of these translations, the sample segments were annotated based on the error typology introduced by (Hansen and Esperança-Rodier, V.2023). The original typology was further categorized for the Chinese - English language pair and inter-rater validation purposes. Specifically, six level-one error types were identified:

- semantic errors (SEM): errors that directly affect the meaning of the text, involving issues like omission, addition, or wrong translation of content/nuance of content;
- logical, structural and cohesion errors (LSC): errors related to the logical flow and coherence of the text, affecting how different parts relate to each other;
- grammatical errors (GRM): errors related to

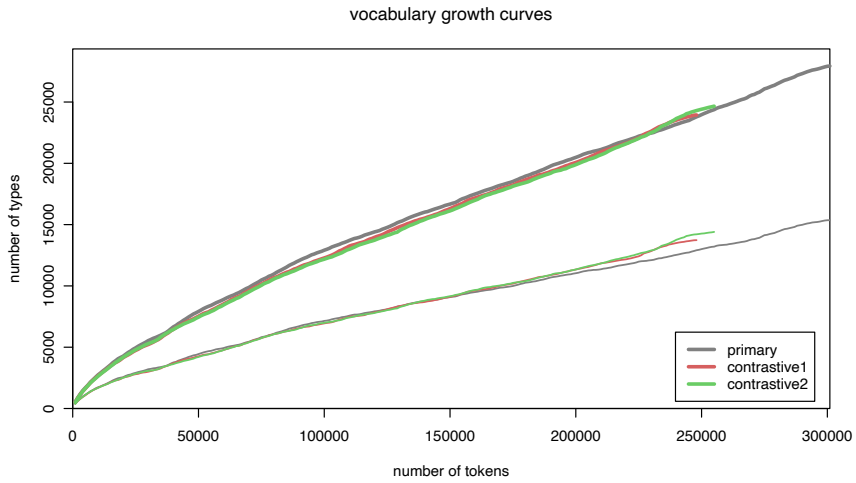


Figure 1: Vocabulary growth curves of our three translations (primary, contrastive1, contrastive2). The lower series of curves corresponds to the hapaxes for primary, ontrastive1 and contrastive2.

the rules of language such as gender, number, tense, and person etc.;

- stylistic errors (STY): errors regarding the style, tone, and appropriateness of the language used;
- stuttering (STU): words repeated for no apparent reason by the MT system;
- non-translation (NTR): source term left untranslated in the target.

Each level one error type contains specific level two and sometimes level three error types. The complete error typology tailored for this task can be found in the appendix.

We use Fleiss' kappa to measure the Level 1 error type inter-rater agreement, and the overall Fleiss' kappa score is 0.288, which can be interpreted as "Moderate agreement" according to (Landis and Koch, 1977)'s classifications. . Fleiss' kappa of Level 1 subgenre annotations is presented in Table 2.

Among all annotated segments, 30.58% segments are considered error-free. 47.71% of them belong to the SEM error type, with the remainder of 11.31% on STY, 4.89% on LSC, and 3.98% on GRM.

4.3.3 Prominent Error Types

Understanding the text in its original language is the basis for literary translation, which requires multi-faceted considerations pinned by context, literary style and cultural nuance. The fine-tuned system attempts to address the greater-than-sentence-level textual features. However, human annotation results have shown that it continues to struggle with contextual analysis, which leads to prominent errors such as non-translation, mistranslation and inconsistent translation or reference of proper nouns and terms, mistranslation of idioms, etc.

Subgenre	**Score** ↑
Modern Military	0.534
Science Fiction	0.344
Ancient Romance	0.321
Fantasy	0.283
Modern Romance	0.283
Thriller	0.152
Online Game	0.143

Table 2: Fleiss' kappa of subgenres. $\kappa = 1$ is perfect concordance, $\kappa = 0$ is no concordance between annotators.

Transliteration is the main way of addressing the character names from Chinese into English (in this case, standard Pinyin is used). Surprisingly, the system failed to maintain consistency of reference to name entities, for example, "宋扶" (song fu) was translated as "Song Fu", "Song Fudge" and "Song Yidao" at places. The character "宋扶" is also mentioned as "宋师弟" or "宋师兄", which were translated literally (see examples in table 3, hand-annotated in bold). Given the nature of fantasy (xianxia) novels, "师兄" (senior brother) or "师弟" (junior brother) is a common way of addressing

王子法一脸惊讶道：“师兄此话怎讲？”	"What do you mean, senior brother?" Prince Charming asked in surprise.
郑金龙笑眯眯道：“师弟，你是在跟我装糊涂吗？宋师弟的死，你们不准备给师门一个交代？”	"Junior brother, are you playing dumb?" Zheng Jin Long said with a smile.“You don’t want to give your sect an account for **Junior brother Song**’s death?"
王子法面容一肃，沉声道：“宋师兄差点坏我蓝玉门好事，宋扶该死！再给我们一次机会，我们还是会这样做！”	"**Senior Brother Song** almost ruined our Lanyu Sect’s business. **Song Fudge** deserves to die! Give us another chance, and we’ll still do this!" the prince said solemnly.

Table 3: Examples for illustration

people under the same sect. Literal translation in this particular context might reduce textual cohesion and such inconsistent reference might confuse target language readers given the numerous consecutive mentions of “brother” in the text. The same issue was observed in the document-level model (contrastive1) result too.

It is difficult for the system to identify a named entity if the name itself or part of the name can be used as a proper noun. For example, “王子法” (wang zi fa) was mistranslated as “Prince Charming”, which was because the system misidentified the first two Chinese characters “王子” (wang zi, literal meaning: prince) as a named entity.

Other inconsistency regarding proper nouns lies in the formality of presentation, i.e., case error, meaning translation going against previous choices regarding the capitalization of series-specific terms. In fantasy novels, sect names and martial arts techniques are prominent terms. However, the capitalization of these terms was not always consistent.

It is challenging for the current system to capture ideas or emotions in culturally specific expressions. For example, the idiom “天下没有不散的宴席” is translated as “there is no such thing as a banquet in the world”. As a literal translation, it omitted the important part of the idiom “不散的” (literal meaning: non-separable / never-ending), which leads to the failure of conveying its figurative meaning “All good things must come to an end”. On the contrary, it did well in translating “哑巴吃黄连” (literal meaning: a mute person eats bitter melons) as “speechless”. The discrepancy between the translation quality of idioms shows that more culture-specific training data is needed to improve the accuracy and idiomaticity of literary machine translation.

4.4 Sentence- vs. Document-based Training Strategies

An important aspect of the competition was the choice to use full chapters with contextualised successive sentences instead of (more) limited contexts usually retained for translation competitions. This resulted in a much bigger dataset than for more standard competitions (in the vicinity of 400 sentences for biomedical tasks). We submitted 2 models based on a similar architecture: *Contrastive1* and *Contrastive2*.

We used as *Contrastive2* a context-agnostic sentence-level transformer model as in (Vaswani et al., 2017) trained on 10 epochs.

We used as *Contrastive1* an on-context transformer model with the exact same architecture as *Contrastive2* but that adopts sliding windows of 3 concatenated sentences pre-trained on 10 epochs to the sentence-level and trained on 4 epochs with concatenated sentences.

Concatenation of 3to3 implies that the source sentence is concatenated to the two previous sentences using end-of-sentence tokens between each of them. A *sliding windows* is when sliding-KtoK model encodes the source windows sentences x_K^i using the end to sentence tokens <*eos*> and a special token <*S*> used to mark sentence boundaries in the concatenation then decode the translation y_K^i

$$x_K^i = x^{i-K+1}\text{<S>}x^{i-K+2}\text{<S>}...\text{<S>}x^i\text{<eos>}$$

$$y_K^i = y^{i-K+1}\text{<S>}y^{i-K+2}\text{<S>}...\text{<S>}y^i\text{<eos>}$$

Another Contrastive model was trained, but unfortunately too late for the submission, based on (Lupo et al., 2022) it has the same specificity than *Contrastive1* with a context discount of 0.01. Context-discount means that the loss function is defined as :

$$\mathcal{L}_{CD}(x_K^j, y_K^j) = CD \cdot \mathcal{L}_{context} + \mathcal{L}_{current}$$

After the submission period, we continued training our contrastive systems. After 55 epochs

of sentence-level pre-training and 14 epochs of document-level training, the system achieved a BLEU score of 21.46 on Test1 test set.

5 Further Research

5.1 Related Research

This subsection discusses related papers.

For fine-tuning mBART, we replicated the parameters tested by (Lee et al., 2022), namely retraining for three epochs. With the same parameter, (Namdarzadeh et al., 2023) have fine-tuned Persian→English and Persian→French with a single short story but nevertheless observed dramatic improvement for Persian→French translations in terms of elimination of hallucinations, English words and morpho-syntactic correction. We have not tried other multilingual Large Language Models such as mBERT (Wu and Dredze, 2019) (based on BERT), mT5 (Xue et al., 2020), XLM-R (Conneau et al., 2019) based on RobertA or the more recent (and bigger) Bloom model (Scao et al., 2022).

For concatenation Transformer, we used some parameters tested by (Lupo et al., 2022) that translated English→German and English→Russian to observe dramatic improvement on Contrapro set (Müller et al., 2018) and English→Russian set (Voita et al., 2019) although with only a slight improvement in BLEU score.

5.2 Future Research

This first collaboration between several universities and backgrounds has discussed English input and was an opportunity to discuss the findings of the competition on literary data and also our insights into the fine-tuning of mBART50 with literary data. We aim to replicate this analysis on Farsi data, as Farsi is one of the 50 languages of mBART50. As is often the case in competitions, we did not train as much as we expected. For the fine-tuning of mBART, we managed to train for three epochs, which is what we found in previous studies (Lee et al., 2022), but for two other submissions, we were training from scratch and could only manage to train for 10 epochs for constrastive2 (sentence-level) and fine-tune for 4 epochs for contrastive1 (document-level). This impacted our results. Evaluating our BLEU score on Test1, we got 22.31 BLEU score for both primary and contrastive2 meanwhile 19.03 BLEU score for constrative1.

6 Conclusion

This paper presented the MAKE-NMTViz system description for the WMT2023 Literary Shared Task. We participated in the Chinese-to-English task with a model trained at sentence level and at document level. We only used the data provided by the organisers but also analysed the translations produced with mBART50 before our submissions. As we did not receive scores from the organisers of the task, we mostly focused on the qualitative analysis of our translations. We resorted to a typology of translation errors and highlighted prominent error types that remained in our translations.

Limitations

During this translation task, we met one limitation with respect to the document-level translation system. In this case, we did not adapt the system to process in Chinese→English language pair. We employed the same setup described in previous works, where the system was trained for English→Russian, English→German and English→French languages.

Acknowledgements

This paper emanated from research partly supported by the MAKE-NMTVIZ project, funded under the 2022 Grenoble-Swansea Centre for AI Call for Proposals/ GoSCAI - Grenoble-Swansea Joint Centre in Human Centred AI and Data Systems (MIAI@Grenoble Alpes (ANR-19-P3IA-0003)), and by a 2021 research equipment grant from the Scientific Platforms and Equipment Committee (PAPTAN project) under the ANR grant (ANR-18-IDEX-0001, Financement IdEx Université de Paris). Sadaf Mohseni benefitted from a Collège de France /Université Paris Cité PAUSE scholarship and Nicolas Ballier from a CNRS research leave at LLF (Laboratoire de Linguistique Formelle), which are gratefully acknowledged. This work was also supported by the CREMA project (Coreference REsolution into MAchine translation) funded by the French National Research Agency (ANR), contract number ANR-21-CE23-0021-01.

References

R. H. Baayen and Elnaz Shafaei-Bajestan. 2019. *languageR: Analyzing Linguistic Data: A Practical Introduction to Statistics*. R package version 1.5.0.

Laurent Besacier and Lane Schwartz. 2015. Automated Translation of a Literary Work: A Pilot Study. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 114–122. ACL.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

J.L. Fleiss et al. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Ana Guerberof-Arenas and Antonio Toral. 2022. Creativity in Translation: Machine Translation as a Constraint for Literary Texts. *Translation Spaces*, 11(2):184–212.

Damien Hansen and Emmanuelle Esperança-Rodier. V.2023. Human-Adapted MT for Literary Texts: Reality or Fantasy? In *Proceedings of the New Trends in Translation and Technology Conference – NeTTT 2022*, pages 178–190. Incoma Ltd.

Yue Jiang and Jiang Niu. 2022. How are neural machine-translated chinese-to-english short stories constructed and cohered? an exploratory study based on theme-rheme structure. *Lingua*, 273:103318.

Yue Jiang and Biyan Yu. 2017. A Contrastive Study on the Rendition of Adjectival Possessive Pronouns in Pride and Prejudice by Human Translation and Online Machine Translation. *Journal of Xidian University*, 2:147–155.

Dorothy Kenny and Marion Winters. forthcoming. Customization, Personalization and Style in Literary Machine Translation. In Marion Winters, Sharon Deane-Cox, and Ursula Böser, editors, *Translation, Interpreting and Technological Changes: Innovations in Research, Practice and Training*. Bloomsburry.

Waltraud Kolb. 2020. Less room for engagement. *Counterpoint*, 4: 26–27.

Taja Kuzman, Špela Vintar, and Mihael Arčan. 2019. Neural Machine Translation of Literary Texts from English to Slovene. In *Proceedings of the Qualities of Literary Machine Translation*, pages 1–9. EAMT.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

En-Shiun Annie Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Ifeoluwa Adelani, Ruisi Su, and Arya D McCarthy. 2022. Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation? *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL'22)*, pages 58–67.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in neural machine translation.

Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. 2022. Focused concatenation for context-aware neural machine translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 830–842, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Lorenzo Lupo, Marco Dinarelli, and Laurent Besacier. 2023. Encoding sentence position in context-aware neural machine translation with concatenation. In *The Fourth Workshop on Insights from Negative Results in NLP*, pages 33–44.

Evgeny Matusov. 2019. The Challenges of Using Neural Machine Translation for Literature. In *Proceedings of the Qualities of Literary Machine Translation*, pages 10–19. EAMT.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

Behnoosh Namdarzadeh, Sadaf Mohseni, Lichao Zhu, Guillaume Wisniewski, and Nicolas Ballier. 2023. Fine-tuning mbart-50 with french and farsi data to improve the translation of farsi dislocations into english and french. In *Proceedings of Machine Translation Summit XIX: Users Track*, pages 152–162, Virtual. Association for Machine Translation in the Americas.

Antoni Oliver González. 2017. InLéctor: Automatic Creation of Bilingual E-Books. *Tradumàtica*, 15:21–47.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Chung-ling Shih. 2016. Can Machine Translation Declare a New Realm of Service? Online Folktales as a Case Study. *Theory and Practice in Language Studies*, 6(2):252–259.

Kristiina Taivalkoski-Shilov. 2019. Ethical Issues Regarding Machine(-Assisted) Translation of Literary Texts. *Perspectives*, 27(5):689–703.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring Document-Level Literary Machine Translation with Parallel Paragraphs from World Literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902, Abu Dhabi. ACL.

Antonio Toral and Andy Way. 2018. What level of quality can neural machine translation attain on literary text? In Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty, editors, *Translation Quality Assessment: From Principles to Practice*, pages 263–287. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Longyue Wang, Zefeng Du, DongHuai Liu, Deng Cai, Dian Yu, Haiyun Jiang, Yan Wang, Shuming Shi, and Zhaopeng Tu. 2023a. Guofeng: A discourse-aware evaluation benchmark for language understanding, translation and generation.

Longyue Wang, Zefeng Du, Dian Yu, Liting Zhou, Siyou Liu, Yan Gu, Yufeng Ma, Bonnie Webber, Philipp Koehn, Yvette Graham, Andy Wray, Shuming Shi, and Zhaopeng Tu. 2023b. Findings of the wmt 2023 shared task on discourse-level literary translation. proceedings of the eighth conference on machine translation (wmt).

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

A Error Typology

- Semantic Errors: Addition (including Overtranslation); Undertranslation (including Omission); Mistranslation (including Opposite Meaning, Nonsense, and Shift in Meaning); Hallucination; Literal Translation.
- Logical, Structural and Cohesion Errors: Referential Cohesion; Relational Cohesion; Function Words; Logic; Coherence with Previous Volumes; Loss.
- Grammatical Errors: Gender; Number; Tense; Person.
- Stylistic Errors: Language Style; Register; Unfitting Paraphrase; Case; Punctuation; Adaptation; Dialogues.
- Stuttering.
- Non-translation.

DUTNLP System for WMT23 Discourse-Level Literary Translation

Anqi Zhao[1], **Kaiyu Huang**[2], **Hao Yu**[1], **Degen Huang**[*]
[1]Dalian University of Technology, Liaoning, China
[2]Institute for AI Industry Research (AIR), Tsinghua University, Beijing, China
user_zaq@mail.dlut.edu.cn; huangdg@dlut.edu.cn;

Abstract

This paper details the submission from the DUTNLP Lab for the WMT23 Discourse-Level Literary Translation in Chinese to English translation direction under unconstrained conditions. Our primary system aims to harness a large language model with various prompt strategies, allowing for a comprehensive exploration of the potential capabilities of large language models in discourse-level neural machine translation. Moreover, we apply detailed data preprocessing methods to filter bilingual data, which proves to be beneficial. Additionally, we assess a widely used discourse-level machine translation model, G-transformer, using different training strategies. In our experimental results, the method employing large language models achieves a BLEU score of 28.16, whereas the fine-tuned method scores 25.26. These findings indicate that selecting appropriate prompt strategies based on large language models can significantly enhance translation performance compared to traditional model training methods.

1 Introduction

The DUTNLP Lab is actively participating in WMT23 Discourse-Level Literary Translation, focusing on Chinese to English translation direction. As observed, prompting large language models (LLMs) has led to outstanding performance across a range of natural language processing (NLP) tasks (Chowdhery et al., 2022; Goyal et al., 2023; Chung et al., 2022). So our research involves experimenting with various prompts and in-context learning strategies, utilizing large language models. Additionally, we conduct experiments to explore the impact of sentence length and data preprocessing methods on translation results.

Our research is primarily anchored in the gpt-3.5-turbo model (Brown et al., 2020), renowned for its outstanding language generation capabilities spanning various domains, from writing to conversations. This model excels at producing natural and fluent text with simple prompts, making it accessible even to individuals without extensive technical knowledge.

Intriguingly, for crafting effective prompts to stimulate the machine translation capability of the large model, we take inspiration from gpt-3.5-turbo. We actively interact with it to derive prompts that can boost translation performance, resulting in the identification of three candidate translation prompt templates. Our evaluation of these prompts in the discourse-level translation task indicates their overall effectiveness, with minor performance variations.

Recognizing the substantial impact of data quality on translation performance, we employ cleaner development corpora for our main experiments. When utilizing large pre-trained models, we conduct a data filtering process through off-the-shelf tools and manual rule-based approaches. Further details will be seen in Session 2.

Given the inherent randomness and flexibility in translations generated by large models, aligning the output with the source text can be challenging. To tackle this challenge, we develop scripts to identify segments with alignment errors and subsequently apply manual corrections for rectification.

To sum up, our contributions can be outlined as follows:

- We have carefully crafted a prompt that has led to a notable performance of 28.16 BLEU (Papineni et al., 2002) on our dataset. This accomplishment suggests a significant improvement over standard document-level machine translation models, including the G-transformer model (Bao et al., 2021), trained with various strategies.
- We have conducted a series of meticulously controlled experiments to systematically in-

[*]Corresponding authors

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 296–301
December 6–7, 2023.

Figure 1: Prompts advised by gpt-3.5-turbo for machine translation.

vestigate the impact of different prompt strategies, batch sentence quantities, and tokenizer methods on the performance of the gpt-3.5-turbo model when apply to the Chinese-to-English discourse-level translation task.

This paper is structured as follows: Section 2 describes the data pre-processing strategies, followed by the details of our method in Section 3. Section 4 presents the experimental results and analysis, and we draw conclusions in Section 5.

2 Data Processing

Contrary to the conventional fine-tuning approach on large language models, our method utilizes a large pre-trained language model combined with prompts. In other words, Our primary experiment do not require further model training. Therefore, we conduct experiments using only a small portion of the development dataset.

Since the data quality significantly impacts our final translation performance, we adopt both traditional data processing methods and manual rules for filtering. The pre-processing strategies are as follows:

- Extract the discourse-level data from the text data with HTML tags and filter out duplicated sentence pairs.
- Filter out sentences containing illegal and invisible characters, like certain emoji symbols, as they may cause alignment issues.
- Normalize punctuation using Moses scripts (Koehn et al., 2007) for English and Chinese. Chinese text is separately segmented by Jieba tool.
- For Chinese, convert full-width format to half-width format and traditional Chinese characters to simplified ones.

	Translation Prompt
TP1	Translate the following sentences "[Insert text here]" from [SRC] to [TGT].
TP2	These sentences "[Insert text here]" are in [SRC] and can be translated to [TGT] as follows:
TP3	Please provide translations of these sentences "[Insert text here]" into [TGT].

Table 1: Candidate translation prompt.

3 Method

To unlock the full potential of large language models, we introduce an innovative approach by seeking guidance from gpt-3.5-turbo for the creation of effective machine translation prompts (Jiao et al., 2023). Specifically, we pose the following query: 'Provide ten concise prompts or templates that can prompt translation.'

The obtained results are shown in Figure 1. Upon observation, we note that the generated prompts are reasonable and similar. Consequently, we consolidate them into three sets of candidate templates, as illustrated in Table 1, where [SRC] and [TGT] represent the source and target language of translation.

In previous studies concerning discourse-level machine translation, it is evident that factors such as varying discourse lengths (Wang and Cho, 2019; Raffel et al., 2019) and different segmentation granularities (Koehn, 2005; Sennrich et al., 2016) can significantly impact translation performance. Consequently, we design a series of comparative experiments to investigate these aspects. Specifically, we segment the document texts into sizes of k and analyze the effects of different text lengths on machine translation performance in our experimental results.

During the segmentation of document text, our goal is to achieve an equitable distribution of text segments and prevent a situation where only a few isolated sentences remain at the end of a document. To address this, we devise a text segmentation algorithm that preserves the data while also ensuring that the number of text portions between segments is as uniformly distributed as possible. The aim is to minimize variance in sentence counts, as illustrated below.

The main strategy is as follows: for a document containing n lines of text, it undergoes slicing based on a specified size of m lines, where the quotient is denoted as p and the remainder as q. If there is a remainder ($q \neq 0$), it indicates the need to slice the text into $n/p + 1$ segments. This results in a new quotient, k, and a new remainder, t. Consequently, the last t segments are allocated a line count of $k + 1$, while the rest of the segments maintain a line count of k.

In traditional machine translation experiments, it is well-recognized that varying segmentation granularities can significantly influence translation quality, particularly in languages like Chinese where clear word boundaries are often absent (Zhao et al., 2013). Therefore, we conduct additional experiments to assess the impact of segmentation granularity on translation performance. Our experiments involve three different segmentation granularities for model input in both Chinese and English datasets: unsegmented, Chinese segmented using the 'Jieba' tool, and Chinese-English segmented using the 'MOSS' tool.

Finally, we compare the performance of our system with commonly used document-level machine translation models. Detailed findings will be presented in the subsequent section.

Translation Prompt	BLEU
TP1	27.92
TP2	27.19
TP3	27.54

Table 2: The results of three candidate translation prompts.

Split the document into k segments	BLEU
k=5	27.73
k=10	27.92
k=15	27.94
k=20	28.08
k=25	27.88
k=30	N/A

Table 3: The results of TP1 with different segment lengths.

4 Results

4.1 Score Analysis

In the discourse-level translation task, we evaluate the performance of three different candidate prompts, as shown in Table 2. Considering these candidate prompts, TP1 yields the highest BLEU score. Therefore, in the subsequent comparative experiments, we consistently employ TP1 as the foundational prompt.

We initially include additional theme information in TP1 based on a suggestion from gpt-3.5-turbo. The theme is related to novels, and we use it to translate the provided sentences from Chinese to English. Surprisingly, the resulting BLEU score is only 27.02, which is even worse than the three base candidate prompts. Consequently, we decide to remove this additional theme information.

For text fragment segmentation, we do experiment with different values of k, including 5, 10, 15, 20, 25, and 30. However, when we set $k = 30$, we encounter errors due to the input being too lengthy for the model to handle. Therefore, we obtain results for the five groups, as shown in Table 3.

We observe that, with the same prompt, varying the length of text segments indeed has an impact on translation performance. When the number of sentences reaches 30 and the token count exceeds 4,096, the system can no longer perform translation. Conversely, when the text length is relatively short ($k = 5$), the model cannot gather enough informa-

Word segmentation granularity	BLEU
unsegmented	27.88
segmented with jieba	**28.16**
segmented with moss	27.53

Table 4: The results of TP1 with different Word segmentation granularity.

Training strategies	BLEU
exp_randinit	21.21
exp_finetune	24.46
exp_mBART	25.26

Table 5: The results of G-transformer with different training modes.

tion, leading to the lowest translation performance. Conversely, overly long text segments ($k = 25$) also weaken performance of the model, potentially introducing noise. Therefore, we choose $k = 20$ as the base for our experiments.

As shown in Table 4, the granularity of text significantly affects the performance of machine translation. Experimental results demonstrate that unsegmented Chinese and English texts are impacted due to the lack of alignment between words, resulting in a slight reduction in translation effectiveness. However, the 'MOSS' segmentation granularity leads to the worst result. We infer that the word segmentation results are too dispersed, making it challenging for the large language model to precisely integrate contextual information for word translation.

Before the widespread use of effective prompts for large-scale models, fine-tuning on pre-trained language models is a common approach to enhance translation performance in specific domains. Therefore, for the comparison experiments, we select a state-of-the-art (SOTA) model designed for document-level machine translation. G-transformer is a straightforward extension of the standard Transformer architecture (Vaswani et al., 2017), using group tags for attention guiding, and introducing locality assumption as an inductive bias to reduce the hypothesis space of the attention from target to source. And we train the G-transformer model using the training corpus provided in the task. This training process involved random initialization, fine-tuning initialization, and fine-tuning on mBART (Liu et al., 2020). The results of these experiments are presented in Table 5.

Comparing the experimental results, it becomes evident that conducting targeted fine-tuning experiments on large language models can enhance machine translation performance. However, it is important to note that this approach falls significantly short of the effectiveness achieved by using prompts on large language models.

4.2 Discourse Analysis

In the context of a document translation (S, T), Lyu et al. (2021) argues that translation consistency should be maintained at the target end if a lexical word w occurs multiple times (two or more times) at the source end.

Due to constraints on time and resources, we conduct manual discourse-level analysis on a limited amount of text. Specific operations are as follows: First, we use a co-reference identification tool (Gardner et al., 2018) to identify all co-reference chains in the target-side documents. We perform data cleaning to extract multiple entity co-reference chains and then compare whether the entity words in the co-reference chains maintain translation consistency.

An example is provided in Table 6. Given that the three candidate prompts exhibit similar discourse characteristics, we choose the large language model gpt-3.5-turbo with prompt TP1 as an example for our analysis. We also introduce the model fine-tuned on the large model mBART for comparison.

Upon observing the result, we notice that even excellent models like ChatGPT may face challenges in addressing certain issues of discourse consistency and coherence. This could be attributed to the extensive training data and the challenge of ensuring coverage of test datasets. On the other hand, fine-tuning strategies, owing to their training on domain-specific data, result in more targeted translations and facilitate the maintenance of translation consistency. This underscores a demand for higher quality document-level translation and could potentially indicate a direction: the need to capture more contextual dependencies.

5 Conclusion

We have presented our experimental study on gpt-3.5-turbo for machine translation, covering translation prompts and robustness. Through careful observation and analysis of the experimental

Source	Reference	Num	Large model with prompt TP1	Num	Finetune on mBART model	Num
佑哥	Brother Assist	12	You Ge	12	You Ge	12
落落	Luo Luo	13	Lulu Luo Luo Luoluo	6 5 2	Luo Luo	13
七月	July	12	July Qiyue	7 6	July	13
烈烈	Lie Lie	19	Lielie Lie Lie	15 4	Lie Lie	14
榜	list	6	board list	4 4	list	5
无誓之剑	Oathless Sword	12	Wu Shi Zhi Jian Oathless Sword	9 2	Oathless Sword	10
韩家公子	Yang Master Han	16	Han Jia Gongzi Han's young master	2 14	Yang Master Han	16

Table 6: The analysis of discourse phenomenon on different translation models.

results, we have noted that the utilization of the large language model with prompts achieves a significant improvement, nearly 3 points higher than the baseline. It even surpasses the currently widely used mBART+fine-tune approach for discourse-level machine translation. We also attempt to enhance translation performance by incorporating in-context information, but this lead to a negative impact. Our future work may include investigating the impact of historical context on translation results and iterative refinement of translation. Simultaneously, we will focus on the recognition and translation of discourse phenomena for large language models.

Acknowledgements

We sincerely thank all the anonymous reviewers for their insightful comments and suggestions to improve the paper. This work was supported by the National Key Research and Development Program of China (2020AAA0108004) and the Key Research and Development Program of Yunnan Province (Grant No. 202203AA080004).

References

Guangsheng Bao, Yue Zhang, Zhiyang Teng, Boxing Chen, and Weihua Luo. 2021. G-transformer for document-level machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3442–3455, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. *ACL 2018*, page 1.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2023. News summarization and evaluation in the era of gpt-3.

Wenxiang Jiao, Wenxuan Wang, JT Huang, Xing Wang, and ZP Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Xinglin Lyu, Junhui Li, Zhengxian Gong, and Min Zhang. 2021. Encouraging lexical translation consistency for document-level neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3265–3277, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36, Minneapolis, Minnesota. Association for Computational Linguistics.

Hai Zhao, Masao Utiyama, Eiichiro Sumita, and Bao-Liang Lu. 2013. An empirical study on word segmentation for chinese machine translation. In *Computational Linguistics and Intelligent Text Processing*, pages 248–263, Berlin, Heidelberg. Springer Berlin Heidelberg.

HW-TSC's Submissions to the WMT23 Discourse-Level Literary Translation Shared Task

Yuhao Xie, Zongyao Li, Zhanglin Wu, Daimeng Wei, Xiaoyu Chen, Zhiqiang Rao, Shaojun Li, Hengchao Shang, Jiaxin Guo, Lizhi Lei, Hao Yang, Yanfei Jiang
Huawei Translation Service Center, Beijing, China
{xieyuhao2,lizongyao,wuzhanglin2, weidaimeng, chenxiaoyu35, raozhiqiang,lishaojun18,
shanghengchao,guojiaxin1,leilizhi,yanghao30,jiangyanfei}@huawei.com

Abstract

This paper introduces HW-TSC's submission to the WMT23 Discourse-Level Literary Translation shared task. We use standard sentence-level transformer as a baseline, and perform domain adaptation and discourse modeling to enhance discourse-level capabilities. Regarding domain adaptation, we employ Back-Translation, Forward-Translation and Data Diversification. For discourse modeling, we apply strategies such as Multi-resolutional Document-to-Document Translation and TrAining Data Augmentation.

1 Introduction

Transformer architectures (Vaswani et al., 2017) have achieved outstanding performance on sentence-level machine translation tasks, but still have some shortcomings when it comes to discourse-level machine translation. Particularly, for machine translation scenarios that are highly discourse-dependent, such as novel translation and conversation translation, the performance is unsatisfactory.

This paper presents the submission of HW-TSC to the WMT23 Discourse-Level Literary Translation shared task. We utilize an effective data cleaning pipeline summarized in our previous works (Wei et al., 2022; Wu et al., 2022; Yang et al., 2021) to process the training data. We employ Regularized Dropout, Forward Translation, Back Translation, Data Diversification to train a strong baseline. On top of the baseline, we apply strategies including Multi-resolutional doc2doc Translation (MR-doc2doc), TrAining Data Augmentation (TADA) to enhance discourse-level translation capabilities.

The general translation model does not work well in novel translation. We found that the biggest factor affecting the quality of translation is domain adaptation; however, domain adaptation cannot solve the consistency of named entity such as names, addresses, and zero pronoun in novel translation. The consistency needs to be optimized by using strategies such as MR-doc2doc and TADA.

2 Data

2.1 Data Source

We use the same training data as that for the general MT shared task to train a sentence-level baseline. Then We use GuoFeng Webnovel Corpus[1] (Wang et al., 2023) and web-crawled novel data for domain adaptation and discourse-level capability enhancement. The data size is shown in Table 1.

	Bilingual	Source	Target
General MT	25M	50M	50M
GuoFeng Webnovel Corpus	1.9M	-	-
web-crawled novel data	10M	100M	400M

Table 1: Bilingual and monolingual data used for training.

2.2 Data Pre-processing

The data preprocessing pipeline follows our previous work (Wei et al., 2021), including deduplication, XML content processing, langid (Lui and Baldwin, 2012) and fast-align (Dyer et al., 2013) filtering strategies, etc. We will not repeat the details here.

3 System Overview

3.1 Sentence-level baseline

We directly employ the model we trained for the general MT shared task as the sentence-level baseline in this task. The following is the strategy we use to train the sentence-level baseline.

[1] http://www2.statmt.org/wmt23/literary-translation-task.html

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 302–306
December 6–7, 2023.

3.1.1 Regularized Dropout

Regularized Dropout (R-Drop) (Wu et al., 2021) improves performance over standard dropout, especially for recurrent neural networks on tasks with long input sequences. It ensures more consistent regularization while maintaining model uncertainty estimates. The consistent masking also improves training efficiency compared to standard dropout. Overall, Regularized Dropout is an enhanced dropout technique that often outperforms standard dropout.

3.1.2 Data Diversification

Data Diversification (DD) (Nguyen et al., 2020) is a simple but effective strategy to boost neural machine translation (NMT) (Bahdanau et al., 2015) performance. It diversifies the training data by using the predictions of multiple forward and backward models and then merging them with the original dataset on which the final NMT model is trained. This method is more effective than knowledge distillation and dual learning.

3.1.3 Forward Translation

Forward translation (FT) (Abdulmumin, 2021) uses source-side monolingual data to improve model performance. The general procedure of FT involves three steps: (1) randomly sampling a subset from large-scale source monolingual data; (2) using a "teacher" NMT model to translate the subset into the target language, thereby constructing synthetic parallel data; and (3) combining the synthetic and authentic parallel data to train a "student" NMT model.

3.1.4 Back Translation

Augmenting parallel training data with back-translation (BT) (Sennrich et al., 2016; Wei et al., 2023) has been shown effective for improving NMT using target monolingual data. Numerous works have expanded the understanding of BT and investigated various approaches to generate synthetic source sentences. Edunov et al. found that back-translations obtained via sampling or noised beam outputs tend to be more effective than those via beam or greedy search in most scenarios. For optimal joint use with FT, we employ sampling back-translation (ST) (Edunov et al., 2018).

3.1.5 Alternated Training

While synthetic bilingual data has been shown effective for NMT, adding more synthetic data may deteriorate performance as synthetic data inevitably contains noise and errors. To address this issue, alternated training (AT) (Jiao et al., 2021) introduces authentic data as guidance to prevent model training from being disturbed by noisy synthetic data. AT views synthetic and authentic data as two types of different approximations for the authentic data distribution. The key idea is to iteratively alternate between synthetic and authentic data during training until convergence. Authentic data provides guidance to overcome noise in synthetic data. By alternating data types, AT ensures the usage of a large amount of synthetic data while prevents model deterioration from noisy data.

3.1.6 Curriculum Learning

A practical curriculum learning (CL) (Zhang et al., 2019) approach for NMT should address two key issues: ranking training examples by difficulty, and modifying the sampling procedure based on ranking. For ranking, we estimate example difficulty using domain features (Wang et al., 2020). The domain feature is calculated as:

$$q(x,y) = \frac{\log P(y|x;\theta_{in}) - \log P(y|x;\theta_{out})}{|y|} \quad (1)$$

Where θ_{in} is an in-domain NMT model, while θ_{out} is an out-of-domain model. The novel domain is treated as in-domain.

We fine-tune the model on the valid set to get the teacher model and select top 40% of the highest scoring data for finetuning.

3.2 Domain Adaptation

We found that the translation style of novel translation and general domain translation is completely different, so domain adaptation is very important. So we finetune the sentence-level baseline model with bilingual/monolingual novel data. For web-crawled novel data, we use 100M Chinese monolingual data and 400M English monolingual data to construct FT and ST corpus respectively, and use GuoFeng Webnovel Corpus bilingual 1.9M, web-crawled novel data bilingual 10M, finally mix the four parts data together and shuffle them.

3.3 Discourse Modeling

Although the translation quality has improved with domain adaptation, it still unable to solve document-level translation problems such as NE

consistency and zero pronoun translation. MR-doc2doc and TADA need to be used to solve the problem. It is expected to further improve the ability of discourse-level translation on the basis of section 3.2. We employ monolingual and bilingual novel data, and reconstruct them according to the method of discourse-level translation.

3.3.1 Multi-resolutional doc2doc

Multi-resolutional doc2doc (MR-doc2doc) (Sun et al., 2020) is a document-level neural machine translation approach that operates on different granularities of the document. It utilizes both sentence-level and document-level information during translation to improve context modeling and overall translation quality. Specifically, we split each document averagely into kparts multiple times and collect all the sequences together. For example, a document containing eight sentences will be split into two four-sentences segments, four two-sentences segments, and eight single sentence segments. Finally, fifteen sequences are all gathered and fed into sequence-to-sequence training. In this way, the model can acquire the ability to translate long documents since it is assisted by easier, shorter sentences and paragraphs. By doing so, the model can acquire discourse-level translation capabilities.

3.3.2 TrAining Data Augmentation

The key idea of TrAining Data Augmentation (Ailem et al., 2021) is to use tags to mark words or phrases that needs to be constrained in the source sentence during translation. When the model encounters a tagged token in the source, it is biased towards directly copying the expected lexical constraint following the tagged source word into the target output. This allows enforcing lexical constraints without changing the core NMT architecture, simply by using tags in the source. The model learns this copy behavior during training when exposed to tagged source sentences and the expected lexical constraints in the target. Thus, the approach can easily guide NMT to satisfy terminology constraints by just tagging the source sentence appropriately. It provides a simple and efficient way to constrain NMT output lexicons by merely adding tags on the source side. We use this method to ensure consistent translation of named entities (such as person names, location names, etc.) at both inference and training phases.

4 Experiments

4.1 Experiment Settings

We use SacreBLEU (Post, 2018) to measure system performances. The main parameters are as follows: the model is transformer-big with 25 encoder layers and 6 decoder layers. It trained using 8 A100 GPUs, batch size is 8192, parameter update frequency is 1, and learning rate is 5e-4. The number of warmup steps is 4000, and the model is saved every 1000 steps. We adopt dropout, and the rate varies across different training phases. R-Drop is used in model training, and we set λ to 5. We use fairseq (Ott et al., 2019) for training.

4.2 Testing Datasets

4.2.1 Simple Set

Simple Set[2] (Wang et al., 2023) contains unseen chapters in the same web novels as the training data.

4.2.2 Difficult Set

Difficult Set[3] (Wang et al., 2023) contains chapters in different web novels from the training data.

5 Results

As shown in Table 2, each step is fine-tuned based on the model from the previous step. In the Domain Adaptation stage (ST, ST & FT & AT & DD, CL), we observe significant s-BLEU improvement, while d-bleu is also improved. In the Discourse Modeling stage, MR-doc2doc can improve both s-bleu and d-bleu. TADA works well on NE consistency, but does not significantly improve d-BLEU and leads to a slight decrease in s-bleu.

As shown in Table 3, We extracted 75 NEs by W2NER (Li et al., 2022) from the test set, which occurred 1241 times in total. We count the word frequency of consecutive and identical NEs as an indicator to evaluate the consistency. We found that the TADA strategy can bring significant improvements in NE consistency.

6 Conclusion

It was believed that algorithm enhancement can make the model handle long inputs so that discourse-level translation would be improved. However, it can only achieve slight improvement.

[2] http://www2.statmt.org/wmt23/literary-translation-task.html

[3] http://www2.statmt.org/wmt23/literary-translation-task.html

	Simple		Diffcult	
	s-BLEU	d-BLEU	s-BLEU	d-BLEU
sentence-level baseline + R-Drop	26.63	15.87	23.47	12.97
+ ST	29.36	22.63	26.02	17.74
+ ST & FT & AT & DD	29.49	26.52	25.97	21.54
+ CL	30.96	26.52	27.4	21.92
+ MR-doc2doc	30.71	26.99	27.27	22.16
+ TADA	30.58	**27.27**	27.12	**22.48**

Table 2: BLEU scores of zh→en NMT system on WMT23 web fiction test set.

models	NE consistency accuracy
sentence-level baseline	43.3%
MR-doc2doc	67.0%
TADA	71.8 %

Table 3: NE consistency accuracy of zh→en NMT system on WMT23 web fiction test set.

It is more important to achieve domain adaptation first for the sentence-level model. The discourse-level translation strategy can get the best performance based on domain adaptation.

References

Idris Abdulmumin. 2021. Enhanced back-translation for low resource neural machine translation using self-training. In *Information and Communication Technology and Applications: Third International Conference, ICTA 2020, Minna, Nigeria, November 24–27, 2020, Revised Selected Papers*, volume 1350, page 355. Springer Nature.

Melissa Ailem, Jingshu Liu, and Raheel Qader. 2021. Encouraging neural machine translation to satisfy terminology constraints. *Le Centre pour la Communication Scientifique Directe - HAL - Diderot,Le Centre pour la Communication Scientifique Directe - HAL - Diderot*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, page 489. Association for Computational Linguistics.

Rui Jiao, Zonghan Yang, Maosong Sun, and Yang Liu. 2021. Alternated training with synthetic and authentic data for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1828–1834.

Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022. Unified named entity recognition as word-word relation classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, page 10965–10973.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30.

Xuan-Phi Nguyen, Shafiq Joty, Wu Kui, and Ai Ti Aw. 2020. Data diversification: a simple strategy for neural machine translation. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 10018–10029.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, page 186. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96. Association for Computational Linguistics (ACL).

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2020. Capturing longer context for document-level neural machine translation: A multi-resolutional approach. *Cornell University - arXiv,Cornell University - arXiv*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, AidanN. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Neural Information Processing Systems,Neural Information Processing Systems*.

Longyue Wang, Zefeng Du, DongHuai Liu, Deng Cai, Dian Yu, Haiyun Jiang, Yan Wang, Shuming Shi, and Zhaopeng Tu. 2023. Guofeng: A discourse-aware evaluation benchmark for language understanding, translation and generation.

Wei Wang, Ye Tian, Jiquan Ngiam, Yinfei Yang, Isaac Caswell, and Zarana Parekh. 2020. Learning a multi-domain curriculum for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7711–7723.

Daimeng Wei, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Xiaoyu Chen, Hengchao Shang, Jiaxin Guo, Minghan Wang, Lizhi Lei, Min Zhang, et al. 2021. Hw-tsc's participation in the wmt 2021 news translation shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 225–231.

Daimeng Wei, Zhiqiang Rao, Zhanglin Wu, Shaojun Li, Yuanchang Luo, Yuhao Xie, Xiaoyu Chen, Hengchao Shang, Zongyao Li, Zhengzhe Yu, et al. 2022. Hw-tsc's submissions to the wmt 2022 general machine translation shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 403–410.

Daimeng Wei, Zhanglin Wu, Hengchao Shang, Zongyao Li, Minghan Wang, Jiaxin Guo, Xiaoyu Chen, Zhengzhe Yu, and Hao Yang. 2023. Text style transfer back-translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 7944–7959, Toronto, Canada.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34:10890–10905.

Zhanglin Wu, Jinlong Yang, Zhiqiang Rao, Zhengzhe Yu, Daimeng Wei, Xiaoyu Chen, Zongyao Li, Hengchao Shang, Shaojun Li, Ming Zhu, et al. 2022. Hw-tsc translation systems for the wmt22 biomedical translation task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 936–942.

Hao Yang, Zhanglin Wu, Zhengzhe Yu, Xiaoyu Chen, Daimeng Wei, Zongyao Li, Hengchao Shang, Minghan Wang, Jiaxin Guo, Lizhi Lei, et al. 2021. Hw-tsc's submissions to the wmt21 biomedical translation task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 879–884.

Xuan Zhang, Pamela Shapiro, Gaurav Kumar, Paul McNamee, Marine Carpuat, and Kevin Duh. 2019. Curriculum learning for domain adaptation in neural machine translation. In *Proceedings of NAACL-HLT*, pages 1903–1915.

TJUNLP:System Description for the WMT23 Literary Task in Chinese to English Translation Direction

Shaolin Zhu and Deyi xiong*
College of Intelligence and Computing, Tianjin University, Tianjin, China
{zhushaolin, dyxiong}@tju.edu.cn

Abstract

This paper introduces the overall situation of the Natural Language Processing Laboratory of Tianjin University participating in the WMT23 machine translation evaluation task from Chinese to English. For this evaluation, the base model used is a Transformer based on a Mixture of Experts (MOE) model. During the model's construction and training, a basic dense model based on Transformer is first trained on the training set. Then, this model is used to initialize the MOE-based translation model, which is further trained on the training corpus. Since the training dataset provided for this translation task is relatively small, to better utilize sparse models to enhance translation, we employed a data augmentation technique for alignment. Experimental results show that this method can effectively improve neural machine translation performance.

1 Introduction

Machine translation, as a core branch of natural language processing, has experienced significant development and received widespread attention in the past few years. Propelled by deep learning and neural networks, architectures like the Transformer(Vaswani et al., 2017) and its derivative models, such as BERT (Devlin et al., 2019) and GPT (Brown et al., 2020), have become mainstream methods for achieving efficient machine translation. These models, by learning underlying representations of language, are able to capture complex relationships and rich semantic information between texts.

Although neural machine translation with dense models has a promising future, it still faces many challenges. One of the main issues with the standard Transformer-based dense multilingual neural machine translation model is the model's capacity bottleneck(Zhu et al., 2021; Fedus et al., 2022b; Cheng et al., 2021). While increasing the model's depth and breadth can effectively enhance its capacity, it severely reduces the model's execution efficiency and increases the hardware requirements for training the model. This often results in the need for large GPU devices, limiting the model's applications. Therefore, in recent years, multilingual neural machine translation based on Mixture-of-Experts (MOE) (Fedus et al., 2022a) has been proposed. Compared to dense models, MOE-based multilingual machine translation activates only a portion of the network parameters during model training and inference (Lepikhin et al., 2021), giving it excellent computational efficiency. Under the same hardware conditions, it can achieve greater model capacity (compared to dense models, capacity can be increased by several tens of times) (Shazeer et al., 2017) and shorter computation time. Therefore, in this translation task evaluation, our basic model framework is based on the MOE Transformer. Furthermore, when there is limited available data, overfitting can easily occur (Wang et al., 2022; Pan et al., 2021). Combining the knowledge of multiple experts can often provide more accurate predictions than a single model. During model training, by allocating experts to focus on different input subsets, MOE can help alleviate the overfitting issue (Szymanski and Lemmon, 1993).

In this paper, we primarily focus on the WMT23 Chinese to English machine translation task. To enhance the model's capacity while maintaining a high computational efficiency, we employ a neural machine translation model based on the MOE Transformer framework. This model can effectively expand the model parameters. Moreover, since it's a domain-specific translation task with limited translation data corpus, we employed a strategy to initialize MOE using dense models effectively. The rest of this paper is organized as follows. In Section 2, we will present the models and methods we designed. Section 3 primarily

*Corresponding author.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 307–311

showcases the experimental results and discusses and analyzes the outcomes. Section 4 concludes the paper and provides an outlook.

2 Method Description

To evaluate machine translation from Chinese to English, we need to construct a machine translation model. Therefore, in section 2.1, we first introduce the model's design and initialization strategy. In section 2.2, we primarily discuss the data alignment augmentation method, aiming to further utilize the data to enhance the model's performance. Finally, we introduce the model's training strategy.

2.1 Model Design

Compared to the MOE model, dense models perform better in bilingual settings (Costa-jussà et al., 2022). Given that the WMT23 machine translation evaluation task has relatively limited corpora, in order to enhance the model's performance, we first pretrain a dense model. Then, we use this dense model to initialize the MOE model. The framework of the model is illustrated in the Figure 1.

We first employ a 6x6 Transformer-based encoder-decoder framework to train the dense model. We can then use the parameters of this pretrained model to initialize the MOE-based translation model. The difference between the dense model and the MoE model lies in the fact that some FFN layers are replaced with MoE layers (Lin et al., 2020), while the rest of the structure remains identical. Therefore, we can directly initialize the embedding, Self-Attention, and Cross-Attention using the dense model. As for the MoE layer, it has a routing module and multiple FFN layers of the same size. We take the FFN layer parameters from the corresponding layer in the dense model (Komatsuzaki et al., 2023), add noise to increase the diversity of the initializing parameters, and then use these noisy FFN layer parameters to initialize each FFN in the MoE layer one by one. For the routing module, we initialize it randomly.

Specifically, our model in this paper adopts three stages. First, we train a basic multilingual neural machine translation model using the Transformer model. Upon successfully training the multilingual machine translation model, we select all of its parameters to initialize the MoE model. We need to create multiple expert sub-networks, and each expert sub-network will replicate the parameters of the corresponding FFN layer.

Next, we use the MoE model for self-supervised learning. Self-supervised learning is an unsupervised learning method that generates its own labels. For the machine translation task, one method of self-supervised learning is to use the original language text as input and then predict its translation. We mask 35% of the input text at random on a per-line basis. Ultimately, we compare the predicted text with the original text, compute the loss, and then update the model parameters. Finally, we use the MoE model, which has undergone self-supervised learning, to initialize a new MoE machine translation model. The expert sub-networks of the new model will replicate the parameters of the self-supervised learning model (Koishekenov et al., 2023). We then continue to train the new model until it meets our performance criteria.

2.2 Training Strategy

We preprocess all the data, removing special characters and standardizing punctuation marks. We uniformly apply SentencePiece (spm) (Kudo and Richardson, 2018) tokenization and construct a unified vocabulary with a size of 32,000. Additionally, we use the fairseq tool (Ott et al., 2019) for binarization. During training and decoding, the vocabulary is shared. We chose the Transformer as the foundational architecture and made improvements upon it to train bilingual models, multilingual dense models, and multilingual MOE models. We uniformly divided the data into training and validation sets. Since there is no test set, the final results are evaluated on the validation set. The model employs Adam (Kingma and Ba, 2015) as the optimizer to update model parameters. Every 30k steps, the model's performance is evaluated using the validation set. We use Polynomial Decay to dynamically adjust the learning rate, with the basic idea being to gradually decrease the learning rate as training progresses. For the dense model, it is trained for 100k steps. For the self-supervised model, we initialize the MOE model parameters using the dense model. We set the number of experts to 32, frequency to 4, expert capacity size to 1.0, and train for 50k steps. For the MOE model, we initialize the MOE model parameters using the self-supervised model. We set the number of experts to 32, frequency to 4, expert capacity size to 1.0, and train for 70k steps. During decoding, we adopt the beam search strategy, and the evaluation metric used is sacrebleu (Post, 2018)

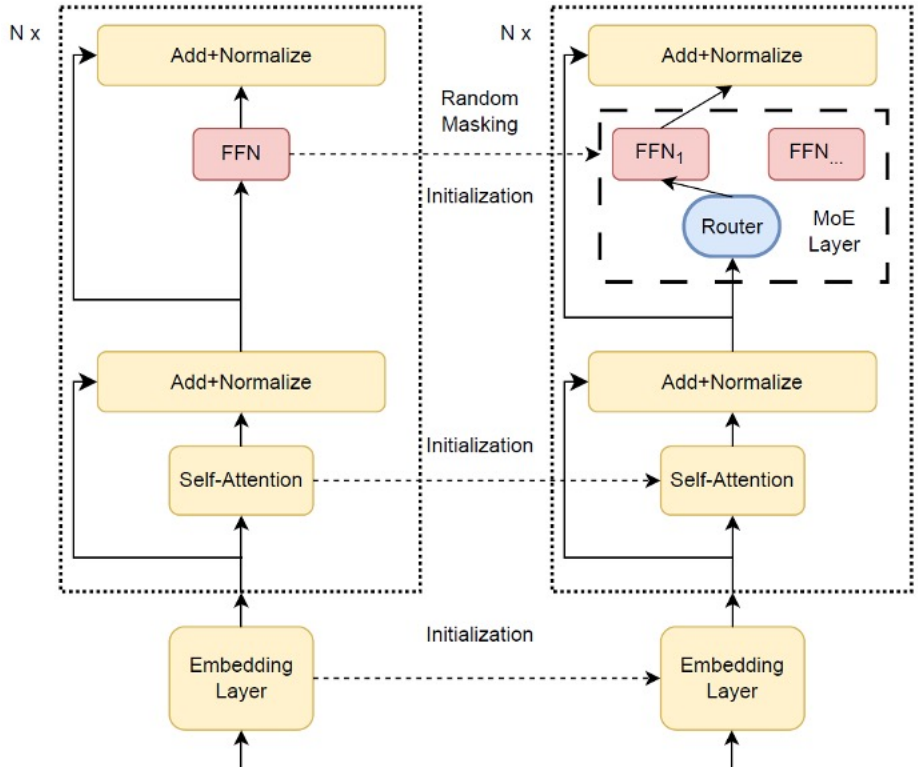


Figure 1: Taking the encoder as an example, the initialization process from the pretrained dense model to MOE is described. The process for the decoder is the same.

3 Experimental Results and Analysis

We first introduce the parameter settings of our trained models, and then analyze the experimental results.

Model	Dense-MOE	Dense
Test1	21.59	19.08
Test2	17.89	15.48

Table 1: Evaluation Results for Dense-MOE and Dense.

3.1 Setting

The model is improved upon fairseq (MoE version) [1]. The training precision is uniformly set to fp16. Both the encoder and decoder are set to 6 layers with 8 attention heads each. The word embedding size is 512, and the hidden layer size is 1024. The loss function used is the cross-entropy function, and the optimizer is Adam, with beta1 set to 0.9 and beta2 set to 0.98. During the pretraining phase, the learning rate is set to 2e-4. A polynomial learning rate scheduling strategy is employed to optimize the learning rate, with warmup set to 4000. Dropout is set to 0.1. Each batch has a maximum of 4096 tokens, and gradients are updated every 4 accumulated batches.

3.2 Experimental Results

For this evaluation task, we did not compare our system with the current state-of-the-art NMT systems. The reason is that the organizers fixed the training data and system configurations to ensure a fair comparison among all participants.We use the Test1 and Test2 provided by the organizers as evaluation targets.

In the experiments, we used sacrebleu as the evaluation metric. From Table 1, we can first observe that the method we employed in this paper achieved better performance compared to the dense model. After training, the dense model has already learned the basic patterns of the dataset. Using these parameters to initialize the MOE model allows the MOE model to start from a more optimal initial state, thereby converging quickly. Using the parameters of the dense model as initial values ensures that the MOE model has already grasped the basic features of the data at the onset of training. This provides a stable starting point for the MOE model, reducing the risks of instability and overfitting during training. Each expert in the MOE model can specifically handle certain distinct patterns or features in the

[1] https://github.com/facebookresearch/fairseq/tree/moe

data. By utilizing the pretrained dense model parameters, each expert in the MOE model can more rapidly identify its area of expertise, leading to a more efficient decomposition of model tasks. Even on the same dataset, due to its structural characteristics, the MOE model can capture more complex patterns in the data. With the initialization from the dense model's parameters, the MOE model can further optimize on this foundation, enhancing the model's expressive capability.

4 Conclusion

This paper introduces the main techniques and methods used for the WMT23 Chinese to English neural machine translation evaluation task. We employ a multilingual neural machine translation model based on the MOE Transformer framework. This model effectively achieves a vast and efficient parameterization. Moreover, given that it's a domain-specific translation task with limited translation data corpus, we utilized an effective strategy of initializing the MOE model using a dense model. This ensures that the MOE model has already grasped the fundamental features of the data at the start of training, providing a stable foundation for the MOE model and reducing the risks of instability and overfitting during training. Experimental results demonstrate that these methods can significantly improve the translation quality of neural machine translation.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Yong Cheng, Wei Wang, Lu Jiang, and Wolfgang Macherey. 2021. Self-supervised and supervised joint training for resource-rich machine translation. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 1825–1835. PMLR.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

William Fedus, Jeff Dean, and Barret Zoph. 2022a. A review of sparse expert models in deep learning. *CoRR*, abs/2209.01667.

William Fedus, Barret Zoph, and Noam Shazeer. 2022b. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.*, 23:120:1–120:39.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Yeskendir Koishekenov, Alexandre Berard, and Vassilina Nikoulina. 2023. Memory-efficient NLLB-200: language-specific expert pruning of a massively multilingual machine translation model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3567–3585. Association for Computational Linguistics.

Aran Komatsuzaki, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. 2023. Sparse upcycling: Training mixture-of-experts from dense checkpoints. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical*

Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018, pages 66–71. Association for Computational Linguistics.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2021. Gshard: Scaling giant models with conditional computation and automatic sharding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li. 2020. Pre-training multilingual neural machine translation by leveraging alignment information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2649–2663. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Demonstrations*, pages 48–53. Association for Computational Linguistics.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. *CoRR*, abs/2105.09501.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Peter T. Szymanski and Michael D. Lemmon. 1993. Adaptive mixtures of local experts are source coding solutions. In *Proceedings of International Conference on Neural Networks (ICNN'88), San Francisco, CA, USA, March 28 - April 1, 1993*, pages 1391–1396. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Wenxuan Wang, Wenxiang Jiao, Yongchang Hao, Xing Wang, Shuming Shi, Zhaopeng Tu, and Michael R. Lyu. 2022. Understanding and improving sequence-to-sequence pretraining for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2591–2600. Association for Computational Linguistics.

Yaoming Zhu, Jiangtao Feng, Chengqi Zhao, Mingxuan Wang, and Lei Li. 2021. Counter-interference adapter for multilingual machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 2812–2823. Association for Computational Linguistics.

Machine Translation for Nko: Tools, Corpora and Baseline Results

Moussa Koulako Bala Doumbouya [*S,F] **Baba Mamadi Diané** [N] **Solo Farabado Cissé** [N]
Djibrila Diané [N] **Abdoulaye Sow** [F] **Séré Moussa Doumbouya** [F]
Daouda Bangoura [F] **Fodé Moriba Bayo** [F] **Ibrahima Sory 2. Condé** [K]
Kalo Mory Diané [N] **Chris Piech** [S] **Christopher Manning** [S]

[S] Computer Science Department, Stanford University. 450 Jane Stanford Way, Stanford, CA 94305
[N] Nko USA Inc. 365 E 169th St. Bronx, NY, US 10456
[F] Friasoft. 9C5M+33, Fria, Guinea.
[K] Kofi Annan University. J986+7P Conakry, Guinea

Abstract

Currently, there is no usable machine translation system for Nko [1], a language spoken by tens of millions of people across multiple West African countries, which holds significant cultural and educational value. To address this issue, we present a set of tools, resources, and baseline results aimed towards the development of usable machine translation systems for Nko and other languages that do not currently have sufficiently large parallel text corpora available. (1) Fria||el: A novel collaborative parallel text curation software that incorporates quality control through copyedit-based workflows. (2) Expansion of the FLoRes-200 and NLLB-Seed corpora with 2,009 and 6,193 high-quality Nko translations in parallel with 204 and 40 other languages. (3) nicolingua-0005: A collection of trilingual and bilingual corpora with 130,850 parallel segments and monolingual corpora containing over 3 million Nko words. (4) Baseline bilingual and multilingual neural machine translation results with the best model scoring 30.83 English-Nko chrF++ on FLoRes-devtest.

1 Introduction

The Manding languages, including Bambara, Maninka, Mandinka, Dyula, and several others, are generally mutually intelligible and spoken by over 40 million people across West African countries including Mali, Guinea, Ivory Coast, Gambia, Burkina Faso, Sierra Leone, Senegal, Liberia, and Guinea-Bissau. Nko, which means 'I say' in all Manding languages, was developed as both the Manding literary standard language and a writing system by Soulemana Kanté in 1949 for the purpose of sustaining the strong oral tradition of Manding languages (Niane, 1974; Conde, 2017; Eberhard et al., 2023).[2] Nko thus serves a role for the Manding languages somewhat akin to Modern Standard Arabic for Arabic languages. It adequately transcribes their essential features such as vowel length, nasalization, and tone (Oyler, 2002; Conde, 2017; Donaldson, 2017) and enables the development of a shared literature.

Since its invention, the use of Nko has been growing. It is taught by literacy promotion associations, and used in newspapers, social media, and electronic communication (RFI, 2016; Rosenberg, 2011; Donaldson, 2019; Diane, 2022). Given that students learn best in their native language (Soh et al., 2021), Nko is particularly valuable for elementary native language education. Unfortunately, Nko and more generally West African languages remain marginalized in West African academic institutions (Kotey, 1975; Bryant, 2020). As a result, and despite the efforts of its courageous community, few academic resources are available in Nko.

Amongst numerous other benefits, computer-assisted translation could be used to facilitate the translation of academic content between Nko and other languages and facilitate projects such as Nko Wikipedia, which currently contains less than two thousand articles, in contrast with French and English Wikipedia with over 2 and 6 million articles respectively (Wikimedia, 2023). Unfortunately, to date, there isn't any usable machine translation (MT) system for Nko, in part due to the unavailability of large text corpora required by state-of-the-art neural machine translation (NMT) algorithms.

Nko is a representative case study of the broader issues that interfere with the goal of universal machine translation. Thousands of languages still don't have available or usable MT systems, mainly due to the unavailability of high-quality parallel text corpora. Recent corpora curation efforts have also resulted in sub-standard data quality for some languages. Some issues reported by (NLLB Team

*moussa@cs.stanford.edu

[1] Also spelled N'Ko, but speakers prefer the name Nko.

[2] ISO-639 code: nqo; ISO-15924 code: Nkoo.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 312–343
December 6–7, 2023.

et al., 2022) and others that we address in this work (see Section 3.3, and 3.6) could have been avoided with the use of an adequate parallel text corpus curation system, which did not previously exist.

This work aims to bootstrap the development of MT systems for Nko and, in the process, to contribute open-sourced resources and tools applicable to other languages. Our main contributions include:

Novel Parallel Text Curation Software. Our first contribution is Fria||el (pronounced Friallel), a cloud-based collaborative parallel text curation software that helps human translators orchestrate copy-editing processes resulting in high-quality corpora. Fria||el is presented in Section 2.

Extension of FLoRes-200 and NLLB-SEED. Our second contribution is the extension of FLoRes-200 and a multilingually aligned version of the NLLB-SEED (NLLB Team et al., 2022) corpora with high-quality Nko translations performed by Nko native speaker experts. Both FLoRes-200 and NLLB-SEED match our educational objective fairly well. Both are built over sentences drawn from Wikipedia, with NLLB-SEED, in particular, covering various fields of human knowledge and activity. They are therefore more diverse than other common parallel texts, such as religious texts.

Language Resource from the Nko Community. Our third contribution is the *nicolingua-0005* corpus, a collection of mono-, bi-, and trilingual corpora curated from data files donated by Baba Mamadi Diané, Solo Farabado Cissé, Djibrila Diané, Nafadji Sory Condé, and Kalo Mory Diané.

Baseline Machine Translation Results. Our fourth and last contribution consists of baseline NMT experiments from English, French, and Bambara transcribed in Latin script to Nko and vice versa. We present bilingual and multilingual transformer-based NMT systems (Vaswani et al., 2017) built using the fairseq toolkit (Ott et al., 2019). At present, results remain quite modest, with the best *eng_Latn* → *nqo_Nkoo* system scoring 30.83 chrF++ on FLoRes-devtest.

All presented software and tools have been publicly released to facilitate further progress on machine translation for Nko and other languages.[3]

[3]Corpora and software on https://github.com/:
common-parallel-corpora/friallel
common-parallel-corpora/common-parallel-corpora
mdoumbouya/nicolingua-0005-nqo-nmt-resources
mdoumbouya/nko-nmt-wmt-2023

2 Fria||el: Collaborative Parallel Corpus Curation System

Recent efforts on collecting multilingual parallel corpora involved sets of data file exchange between various translation teams (Federmann et al., 2022). This process is error-prone as it doesn't allow the systematic tracking of individual corpus entries through a curation quality process. Other recent similar efforts such as NLLB-SEED (NLLB Team et al., 2022), unnecessarily resulted in bi-text data rather than the intended multi-text because the reference files ended up being modified and re-ordered by various translation teams (see Section 3.3). Adequate software could have helped avoid such issues.

We propose Fria||el, a collaborative system designed to help distributed translation teams produce large multilingually aligned high-quality parallel text corpora. The system design particularly emphasizes suitability for use in various contexts, supporting web and mobile device usage and use in an offline mode. Its design goals include: itemized curation, automatic work organization, collaborative copyediting, and localization to translators' preferred user-interface language and preferred source languages to translate from (Figure 1).

2.1 Previous Tools and Multilingual Parallel Corpora Creation Processes

Masakhane Similarly to (Nekoto et al., 2020), this work is an effort towards African language technology development. Our work is participatory in the sense that we are a diverse team of computer scientists, linguists, and native speakers of Nko and other West African languages. We expect that our approach, and the parallel text curation software we release with this paper, Fria||el, will be valuable for MT technology development for other languages.

ParaText ParaText (SIL International & United Bible Societies, 2023) is specialized software for Bible translation projects. Its features include team management, task assignments, notes, collaborative document editing, multilingual dictionaries, and various biblical resources. It also allows a side-by-side comparison of biblical passages from various sources or in various languages. Paratext is not suited for general-purpose parallel corpus curation for MT. There is no indication that ParaText or any such software was used in the curation process of recent multilingual parallel corpora such as NLLB-SEED, FLoRes-200, and NTREX-128.

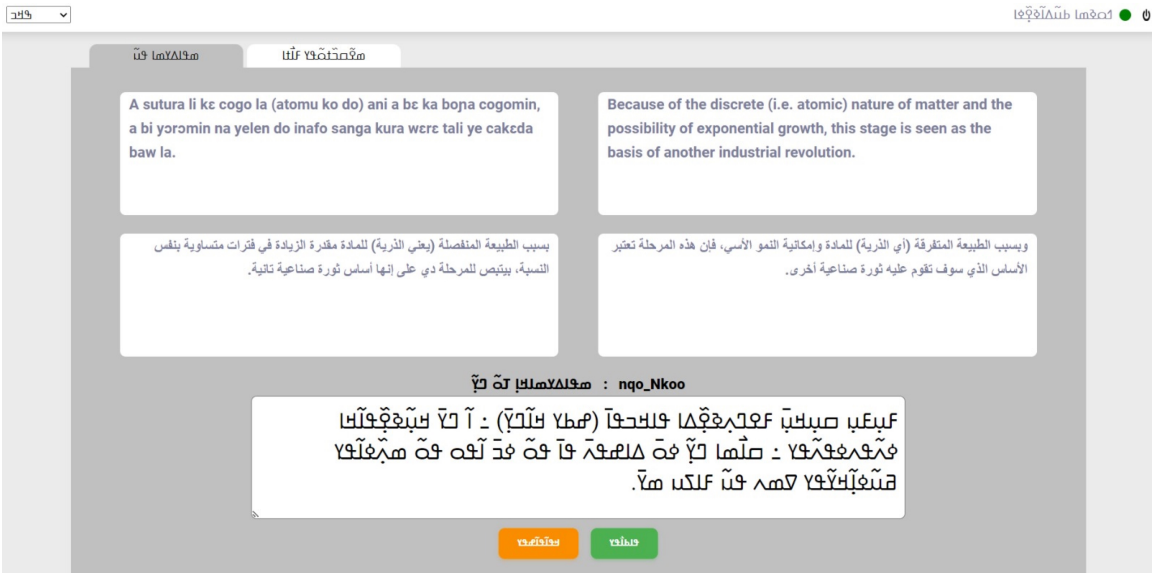


Figure 1: Fria||el's user interface for a Nko translator simultaneously inspecting multiple parallel variants of the same segment from the Multitext-NLLB-SEED corpus. All labels are localized to Nko. The source language fields are also localized to their own language's writing direction: LTR for Bambara in Latin script and English; and RTL for Moroccan Arabic and Egyptian Arabic. The translated text is localized to Nko's writing direction (RTL).

NLLB-SEED and FLoRes-200 The curation process of FLoRes-200 involved teams of translators and reviewers who underwent a vetting process. The QA team reviewed a 20% subset of data files with 3000 entries produced by translation teams. Data files falling below the 90% quality threshold were returned for rework. NLLB-SEED underwent a less rigorous quality control process. The curation process was English-centric. Translators were required to be proficient in English. Translation to the majority of languages was also done from English, with the following exceptions: In NLLB-SEED, Ligurian, was translated from Italian, In FLoRes-200, some Arabic languages were translated from Modern Standard Arabic. As noted by the authors, there are qualified translators who may not speak English, and several languages may be easier to translate from non-English sources.

NTREX-128 NTREX-128 (Federmann et al., 2022) was curated as follows. The English reference file was sent to a translation provider that produced translations. Source-based direct assessment was performed on the translated files by a different provider using the Appraise platform (Federmann, 2018) to generate segment-level quality scores. Segments with a score below a specified threshold were returned for correction. The translation process and quality control method of the translation provider were not specified.

Fria||el is a collaborative parallel text curation software system that tracks individual segments through a translation and copyedit workflow. Each segment is translated by one translator, and subsequently sequentially copyedited by other translators. Fria||el allows translators to simultaneously inspect variants of the source segment in multiple languages. This results in segments translated and copyedited in the context of different subsets of source languages. In addition to the final parallel corpus, Fria||el also yields copyedit logs, which could be valuable in various modeling scenarios.

2.2 Design Goals

Fria||el was designed with the following goals:

Itemized Curation Each corpus segment is individually tracked through the curation process in which it is translated to the target language and subsequently reviewed and copyedited several times.

Automatic Task Assignments Translation and copyediting tasks are automatically assigned to translators with fixed lease periods. Uncompleted tasks are automatically reassigned upon expiration.

Collaborative Copyediting Each segment is translated once and copyedited two or three times, following the workflow in Figure 2. Segments for which the first or second verification results in edits are copyedited a third time. A given translator can only perform a task on a given segment once.

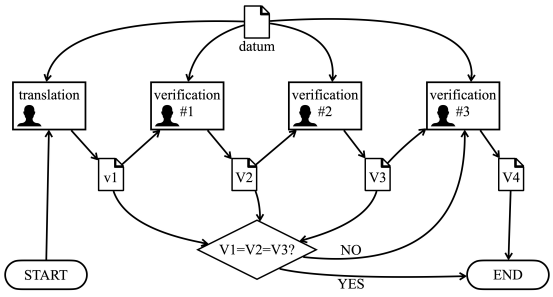


Figure 2: Translation workflow for a multilingual segment (datum). The initial translation (*v1*) is approved or copyedited by two other translators (*v2*) and (*v3*). If any copyediting occurs, a third copyediting task is assigned to a fourth translator who either approves the current translation or performs a final copyedit (*v4*).

Multilingual Sources While performing translation and copyediting tasks, translators can simultaneously inspect segments in several languages configured according to their preferences.

Machine-Generated Sources Datasets can be augmented with additional machine-generated variants of segments such as machine translations, transliterations, and detransliterations.

Responsive Web Design Fria||el is a web application that automatically adapts the layout of its component to the user's screen size. This makes it usable on desktop and laptop computers as well as on mobile phones and tablets.

Resilience to Connectivity Disruptions Translators who temporarily lose their internet connectivity can seamlessly keep working offline on their currently assigned translation and verification tasks. Their work is automatically synchronized with the central database when their connectivity is restored.

Internationalization and Localization Fria||el is internationalized (i18n) in that all user-facing strings are externalized into a translatable resource file, and the writing direction and text alignment of translation source and target languages are configurable. As a result, the user interface is localized (L10n) to the translator's preferred user-interface language, and to each source language (Figure 1).

2.3 Software Components

This section provides details on Fria||el's software components that collectively realize the design goals specified in Section 2.2.

2.3.1 Workflow Manager

Both the Workflow and Task Managers are implemented as Firebase cloud functions that are triggered at fixed time intervals. A workflow entity is inserted for each parallel segment with an initial *active* state. The Workflow Manager periodically inspects workflow entities and (1) creates the next task if needed, and per the workflow management rules, (2) moves the workflow to the *completed* status if all related tasks have been completed and there is no need to create additional tasks or (3) nothing, if the workflow has an uncompleted task.

2.3.2 Task Manager

When triggered, the Task Manager revokes all expired task assignments and assigns unassigned translation and copyedit tasks to users according to their roles. The maximum number of tasks assigned to each user is fixed. A given user is never assigned a task related to a segment on which they have previously completed a task. The Task Manager also ensures that a copyedit task is only assigned to a user with the appropriate verification skill level (*L1*, *L2*, or *L3*) for the first, second, and third copyedit rounds. Each translator account is configured with specific verification skill levels.

2.3.3 Data Model and Storage

Google Firestore, a document-oriented NoSQL database, is used for data storage. The central application database is accessed by data import/export scripts, the WorkflowManager, the Task Manager, and the user interface. It contains the following collections of documents:

datasets: One collection per imported dataset. Each document represents a multilingual segment and contains all available translations of the segment, each annotated with its language and writing system. See Figure 10.

workflows: Each document represents a prioritized workflow entity. The WorkflowManager (Section 2.3.1) periodically inspects workflow entities by priority order and creates task entities as per the workflow management logic.

annotation-tasks: Each document is a task of a specific type (translation or copyedit) related to a specific multilingual segment. Each task has a status (unassigned, assigned, completed). Tasks are assigned to translators by the Task Manager.

users: Each document represents a translator and specifies whether they can be assigned translation (*isActiveTranslator*) and copyedit (*isActiveV-*

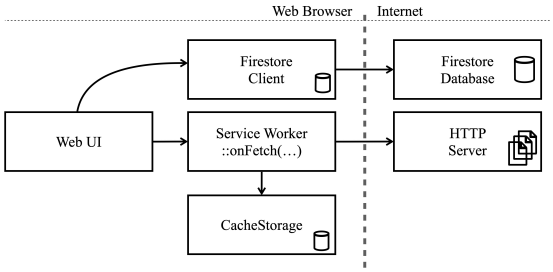


Figure 3: The software uses Firestore's client library's offline mechanism and *cached-aside* HTTP resources to be resilient to intermittent internet disruptions.

erifier) tasks. Translator documents also store the source languages the translator prefers to translate from, subject to availability in the source corpus. User documents also specify a *verifierLevel*, which indicates the maximum copyediting round the translator can participate in for a specific segment.

config: Contains language writing direction configuration. Languages are assumed to be left-to-right unless explicitly marked right-to-left.

2.3.4 User Interface

The user interface is a responsive web application that is usable on a variety of devices, including mobile phones, tablets, desktops, and laptops (Figure 1). It directs authenticated translators to their workspace where they can perform translation (first tab) and copyediting (second tab) tasks that are assigned to them. The task assignment process is transparent to translators. One task is displayed at a time. The prioritized list of tasks assigned to the current translator is kept in a cache for resilience to intermittent internet disruptions. The connection status is indicated by the green circle (top-right).

When performing translation tasks, translators simultaneously inspect the source segment in several languages (top four text fields) and write a translation in the target language text field (bottom). When the 'submit' button (green) is selected, the translation is recorded and the next task is displayed. Translators can also skip the current task by selecting the 'skip' button (orange). When performing copyediting tasks, the bottom text field is initialized with the latest version of the translated segment (Figure 2). The translator may leave the translation intact or copyedit it before submitting.

2.3.5 Offline Mode

The software is a web application designed to be resilient to intermittent internet disruptions. This is achieved with Google Firebase's client library (Google, 2023), which supports offline read and write operations by leveraging a client-side *eventually consistent* (Burckhardt et al., 2014) LRU cache, and *cached-aside* (Pamula et al., 2014) HTTP resources, implemented with two web APIs supported by the majority of web browsers: CacheStorage and ServiceWorker (w3.org, 2022; Mozilla, 2023b,a). After the initial loading of the web application in a web browser, a ServiceWorker is registered to intercept HTTP fetch events. If the remote web server is reachable, the ServiceWorker fetches remote HTTP resources (e.g., HTML, CSS, javascript, image files) and stores them in a CacheStorage before returning them to the caller; otherwise, cached resources are served. The entire process is transparent to the user. See Fig 3.

2.3.6 Translator Copyedit Logs

In addition to the final version of the translated segments, the data Fria||el also outputs their initial translation (*v1*), and the versions of the same entries after the first, second, and third copyediting rounds (*v2*, *v3*, *v4*) – see the workflow in Figure 2. Copyediting logs can be valuable in developing language and machine translation models.

2.3.7 Data Import and Export

Fria||el includes the following administrative Python scripts for importing and exporting parallel corpora and other reports. `load_dataset.py` imports a new parallel corpus from its original data files. Pre-processing may be required to adapt to various original dataset formats. `create_translation_workflows.py` creates active translation workflows for an imported dataset. `system_report.py` displays the number of workflows and tasks by status by dataset. `export_dataset.py` exports translations and translator edits for a curated dataset in a csv file. Post-processing may be required to adapt to a desired format. `accounting_statements.py` generate completed tasks by user by dataset by month. This data can be imported into an accounting system to generate payroll for translators.

2.4 Qualitative User Study

Nko translators used Fria||el to translate FLoRes-200 (dev, devtest) and Multitext-NLLB-SEED to nqo_Nkoo, and to copyedit each segment two or three times. The following sections present an analysis of their responses to a survey questionnaire

(Figure 6). Quantitative measures on their copy-editing logs are also discussed in Section 2.4.6.

2.4.1 Usability

Nko translators praised the simplicity of the user interface. They appreciated the automatically organized itemized copyediting-based data curation process. They highlighted the localization features, particularly, the fact that the user-interface is available in Nko and that the presentation was adequate for both right-to-left and left-to-right source languages and the target language. They valued the offline functionality that allowed them to temporarily continue working without an internet connection. Furthermore, they found the task counters displayed on the user-interface helpful. They noted two usability-related limitations: First, it was not possible to directly go back to a task after submitting it. Second, although the software allowed them to continue working offline, it did not allow them to perform the initial authentication while offline.

2.4.2 Translation Process

Nko translators found the fact that source segments were visible in multiple languages beneficial. They said that the ability to inspect the same segment in multiple languages facilitated its translation to Nko. They also mentioned that the itemized translation tasks, which presented one segment at a time, decreased the likelihood of translation mistakes.

An improvement they requested is the addition of a translation memory including dictionary entries and previously translated expressions.

2.4.3 Copyediting Process

Nko translators found Fria||el's multi-pass copy-editing process effective for finding and correcting translation mistakes. They mentioned that the fact that segments were consecutively assigned to different translators for copyediting led to higher-quality translations as it is easy to overlook one's own mistakes. Because each translator had a different translation source language configuration, Nko segments were translated from and copyedited against their versions in different sets of languages, which Nko translators found enriching.

2.4.4 Mistranslations

Types of mistranslations Nko translators noted during the copyediting process included typos, omitted words, grammatical errors, incorrect word sense translations, incorrect translations of named entities, and punctuation errors. They noted that word sense was sometimes hard to disambiguate without the full context of segments. For instance, the English word *state* maps to different Nko words based on the sense of the word (political community vs. a particular condition of a person, place, or thing). They also noted punctuation errors, particularly the use of the Arabic comma (U+060C) instead of the Nko comma (U+07F8), and spacing around that punctuation. Finally, they reported that translators using different source languages would sometimes disagree on named entity translations.

2.4.5 Disagreements

Nko translators reported few disagreements on language standards. They also reported using existing English-Nko and French-Nko dictionaries for consistency. During the translation of FLoRes-200, NLLB-SEED to Nko, translators participated in weekly team meetings and routinely consulted each other over video conferences and phone calls. They deferred the few cases of disagreement and perplexing questions to the most senior translator.

2.4.6 Copyediting Metrics

Table 1 summarizes the size of the translated corpora in segments and Nko words, as well as the percentage of segments that were edited in each verification round, and the related edit magnitudes, computed as edit distances. The number of edited segments and related edit magnitudes generally decreased as copy-editing rounds progressed.

3 Nko Corpora for Machine Translation

This section discusses the extension of FLoRes-200 and NLLB-SEED to Nko, which included the multilingual alignment of NLLB-SEED, and the use of Fria||el to translate those corpora to Nko. This section also introduces *nicolingua-0005*, a collection of monolingual corpora and bi- and trilingual parallel corpora donated by Nko community members.

3.1 Translation of FLoRes-200 and NLLB-Seed to Nko

Nko native speaker experts Baba Mamadi Diané, Solo Farabado Cissé, and Djibrila Diané, used Fria||el to translate Multitext-NLLB-SEED, FLoRes-dev, and FLoRes-devtest to Nko. They worked from Cairo (Egypt), Banankoro (Guinea), and New York (USA), and with the rest of the team, participated in weekly video conference meetings.

corpus	segments	words	$v1 \rightarrow v2$ edited	$v1 \rightarrow v2$ edit distance	$v2 \rightarrow v3$ edited	$v2 \rightarrow v3$ edit distance	$v3 \rightarrow v4$ edited	$v3 \rightarrow v4$ edit distance
FLoRes-dev	997	27,361	83%	38.75 ± 1.55	67%	50.48 ± 2.10	71%	11.74 ± 0.65
FLoRes-devtest	1,012	29,503	87%	61.74 ± 1.81	93%	9.69 ± 0.64	24%	2.79 ± 0.15
NLLB-SEED	6,193	184,138	48%	45.97 ± 1.11	35%	38.94 ± 1.16	35%	11.96 ± 0.48

Table 1: Percentage of edited Nko segments, and related mean± standard error of edit magnitudes (edit distance) resulting from the translation of FLoRes-dev, FLoRes-devtest, NLLB-SEED to Nko

3.2 Translation Process

The initial translations of FLoRes-dev, which our translators performed using spreadsheets, were imported into Fria||el after our software engineers completed its development. The copyediting tasks for FLoRes-dev, and the translation and copyediting tasks for FLoRes-devtest and Multitext-NLLB-SEED were entirely performed using Fria||el. The system was designed not to allow translators to copyedit their own translations or previous copyedits. This constraint made the proposed translation workflow impossible given the size of our team of translators. As a workaround, an additional user account was created for the two most experienced translators to allow third copyediting rounds.

Each segment was translated once and copyedited two or three times. The resulting curated Nko data files are summarized in Table 2. The multilingually aligned NLLB-SEED dataset (Multitext-NLLB-SEED), FLoRes-dev, and FLoRes-devtest, all extended with Nko translations along with copyedit logs, collectively make up common-parallel-corpora ver. 2023-06-19 summarized in Table 3.

3.3 Multilingual Alignment of NLLB-SEED

The original NLLB-SEED dataset consists of pairwise parallel corpora between English and each other language but suffers from the complication that many of the source English sides are slightly different from each other, variously due to minor copyediting, and reordered and added entries.

Multitext-NLLB-SEED is a multilingually aligned version of NLLB-SEED that fixes this limitation. It was created as follows: A consensus *eng_Latn* reference file was manually edited by human comparison of all existing reference *eng_Latn* files. The lines of each *eng_Latn* file were matched (binary assignment matrix $M_{i,j}$) to the lines of the consensus *eng_Latn* file by minimizing the sum of the edit distances $E_{i,j}$ between matched lines (Equation 1). The optimal

Translations		
lines	words	file
6193	184138	SEED/nqo_Nkoo
997	27361	FLoRes/nqo_Nkoo.dev
1012	29503	FLoRes/nqo_Nkoo.devtest

Translator Edits		
lines	words	file
6193	170555	SEED/nqo_Nkoo.v1
6193	177703	SEED/nqo_Nkoo.v2
6193	182843	SEED/nqo_Nkoo.v3
6193	184138	SEED/nqo_Nkoo.v4
997	24455	FLoRes/nqo_Nkoo.dev.v1
997	25656	FLoRes/nqo_Nkoo.dev.v2
997	26541	FLoRes/nqo_Nkoo.dev.v3
997	27361	FLoRes/nqo_Nkoo.dev.v4
1012	25924	FLoRes/nqo_Nkoo.devtest.v1
1012	27771	FLoRes/nqo_Nkoo.devtest.v2
1012	29521	FLoRes/nqo_Nkoo.devtest.v3
1012	29503	FLoRes/nqo_Nkoo.devtest.v4

Table 2: Extensions of FLoRes-200 (dev, devtest) and Multitext-NLLB-SEED to Nko. The *nqo_Nkoo* data files are parallel with 40 other languages in NLLB-SEED, and 204 other languages in FLoRes-200. FLoRes-test, which is not publicly available, was not translated.

CPC subset	lines	langs.	tr. edits langs.
Multitext-NLLB-SEED	6193	41	1
FLoRes-dev	997	205	1
FLoRes-devtest	1012	205	1

Table 3: Summary of common-parallel-corpora version 2023-06-19. All entries are parallel across all languages. Translator edits are only available for nqo_Nkoo.

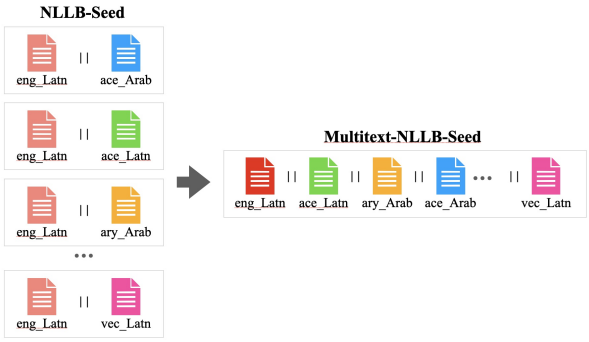


Figure 4: Multitext-NLLB-SEED is a multilingually aligned version of the original NLLB-SEED dataset.

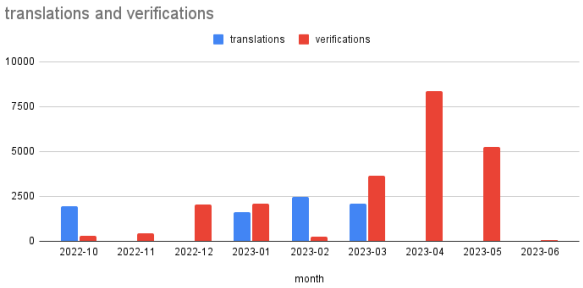


Figure 5: From October 2022 to June 2023, 8,202 translations and 22,426 verifications/edits were performed to produce high-quality translations of FLoRes-200 and Multitext-NLLB-SEED to Nko.

line matching M^*, obtained using the scipy package (Virtanen et al., 2020), was used to re-order each non-English language file to match the order of the consensus *eng_Latn* file. Two unmatched lines from (*eng_Latn*, *kas_Deva*) and one from (*eng_Latn*, *lij_Latn*) were discarded.

$$M^* = \arg\min_{M} \sum_{i,j} M_{i,j} E_{i,j} \quad (1)$$

The resulting re-ordered non-English language files and the consensus eng_Latn file constitute the Multitext-NLLB-SEED corpus, containing 40 parallel language files; see Figure 4. Multitext-NLLB-SEED was loaded in Fria||el in lieu of the original NLLB-SEED corpus, enabling translators to inspect each segment in multiple languages, and resulting in an expanded multilingually aligned corpus.

3.4 Translation Source Languages

Source languages were configured in Fria||el according to the preferences of each translator. Collectively, they translated from *fra_Latn*, *eng_Latn*, *ary_Arab*, *arz_Arab*, and *bam_Latn*. Note that *fra_Latn* is not available in NLLB-SEED. *bam_Nkoo* was detransliterated from *bam_Latn* using a neural detransliterator (Doumbouya, 2022); however, translators did not find this source useful and preferred not to enable it in their configuration.

3.5 Further Notes on Manding languages

Nko was developed as a standardized form of the Manding languages. The aim was a standardized language and writing system, which could serve a similar role to Modern Standard Arabic with respect to various regional Arabic languages. Manding languages, which include Mandinka and Bambara, are a subgroup of the Mande language family and are generally mutually intelligible to speakers. Bambara, written in a Latin script, is currently the best-supported Manding language, available in Google Translate and in NLLB-SEED. Our Nko translators are also fluent in Bambara.

3.6 Quality of bam_Latn in NLLB-SEED

Our Nko translators noted the following quality issues with NLLB-SEED's *bam_Latn* data: (1) The data contains too much French vocabulary not enough Manding vocabulary. (2) Some entries do not match their English counterpart at all. (3) Some entries are entirely in French; examples are shown in Figure 11. (4) The *bam_Latn* data completely lacks tonal marks, which are important in Manding languages (e.g., many nouns are indistinguishable without tonal marks, such as *bird*, *belly*, *inside*; the definite and indefinite inflections of nouns cannot be distinguished without tonal marks (*I saw a person* vs. *I did not see any person*); and nouns that can be used as a verb and their verb form cannot be distinguished (*get out!* vs. *to get out*). *bam_Nkoo*, detransliterated from *bam_Latn* was included in the corpus; however, some Nko translators did not find it useful and preferred to not enable it as a source.

3.7 *nicolingua-0005* Corpus

nicolingua-0005 is curated from files donated by Nko community members for the purpose of developing machine translation for Nko. It is comprised of 3.9 million Nko words with 25K (Nko, English, French) parallel segments, 59K (Nko, English) parallel segments, 45K (Nko, French) parallel segments, and a monolingual corpus of 3.3 million Nko words. Included datasets were curated from files provided by Baba Mamadi Diané, Solo Farabado Cissé, Djibrila Diané, Nafadji Sory

type	languages	segments	nqo words
trilingual	*nqo_Nkoo*, *eng_Latn*, *fra_Latn*	25 848	256 934
bilingual	*nqo_Nkoo*, *eng_Latn*	59 442	283 279
bilingual	*nqo_Nkoo*, *fra_Latn*	45 560	129 789
monolingual	*nqo_Nkoo*	N/A	3 291 371
total		130 850	3 961 373

Table 4: Summary of *nicolingua-0005*

Condé, and Kalo Mory Diané. See Table 4 and Appendix D for more details on the constitution of the corpus. A datasheet questionnaire based on (Costa-jussà et al., 2020) is presented in Appendix E.

4 Baseline Machine Translation Experiments

This section describes Transformer (Vaswani et al., 2017) based encoder-decoder neural machine translation models built using the fairseq toolkit (Ott et al., 2019). Both bilingual and multilingual translation models are explored. At present, results remain quite modest, with the best model achieving a 30.83 *eng_Latn* → *nqo_Nkoo* chrF++ score on the CPC/FLoRes-devtest corpus.

Eight models were trained: The bilingual unidirectional models 200.11 and 200.16, the multilingual model 201.16, and its variant that is trained to also autoencode Nko segments 202.16, and Models 206.19, 207.19, 208.19 and 209.19, which explore three different ways of specifying language tokens.

4.1 Datasets

common-parallel-corpora (CPC) and *nicolingua-0005*, described in Section 3 are used to build baseline NMT models for the following translation directions: *nqo_Nkoo* ⇄ *eng_Latn*, *nqo_Nkoo* ⇄ *fra_Latn*, and *nqo_Nkoo* ⇄ *bam_Latn*. The subsets of those corpora used to train, validate, and test the models are specified in Tables 11 and 12.

4.2 Tokenization

Byte-pair encoding (BPE) (Sennrich et al., 2016) is employed to perform sub-word tokenization. In each training experiment, the BPE model is trained on a token corpus constructed by concatenating all data files containing the languages of interest in the training set. In all cases, the BPE model is trained to produce 15K sub-word units.

4.3 Models

Eight models were trained. The first two, 200.11 and 200.16, are unidirectional bilingual *nqo_Nkoo* ⇄ *eng_Latn* models. The last six, 201.16, 202.16, 206.19, 207.19, 208.19, and 208.19 are multilingual *nqo_Nkoo* ⇄ *eng_Latn*, *nqo_Nkoo* ⇄ *fra_Latn*, and *nqo_Nkoo* ⇄ *bam_Latn* models.

4.3.1 Bilingual Models

200.11 is the baseline bilingual *nqo* ← *eng* model. 200.16 differs from 200.11 in terms of model architecture and hyper-parameters. 200.16 and the multilingual models 201.16 and 202.16 have identical architectures and training hyper-parameters.

Model 200.11 is a Transformer-based (Vaswani et al., 2017) encoder-decoder sequence-to-sequence model consisting of 5 encoder and 5 decoder layers, each with a 512-dimensional token embeddings and 2048-dimensional feed-forward networks, 2 attention heads per layer, and a layer normalization module before each layer. Its architecture and training hyper-parameters are identical to the baseline system of the AmericasNLP 2021 Shared Task on Open Machine Translation (Mager et al., 2021), except for the following differences: (1) encoder and decoder embeddings are not shared, (2) Subword Regularization (Kudo, 2018) and BPE-dropout (Provilkov et al., 2020) are not employed in BPE tokenizer training, (3) larger batches are employed during training, (4) gradient clipping is applied during training.

Model 200.16 This model is only different from 200.11 in that it is deeper (6 encoder layers and 6 decoder layers), and that it is trained with a higher token dropout probability (0.6 instead of 0.4).

4.3.2 Multilingual Models

Our multilingual models are trained on parallel corpora obtained by concatenating all available (*nqo* ⇄ *eng*, *nqo* ⇄ *fra*, *nqo* ⇄ *bam*) bitext and prefixing the source segments with language tokens as introduced by (Johnson et al., 2017). Similarly to (Wicks and Duh, 2022), models 206.19, 207.19, 208.19, and 209.19 compare the effect of various approaches to constructing source-side prefixes.

Model 201.16 is the baseline multilingual model. It has the same architecture and training hyperparameters as the bilingual model 200.16, but it is trained on multilingual data and it employs target language token prefixes (Table 5).

Model 202.16 employs target language tokens just like 201.16, but its training set also contains *nqo* → *nqo* pairs where each side is the same sentence from monolingual Nko corpora in *nicolingua-*

model	prefix
200.11 200.16	(none)
201.16 202.16 206.19	<to_tgt_Lang>
207.19	<from_src_Lang> <to_tgt_Lang>
208.19	<from> <src_Lang> <to> <tgt_Lang>
209.19	<from_src_Lang_to_tgt_Lang>

Table 5: Specification of source sequence language token prefixes used in our multilingual translation models.

0005. Consequently, 202.16 performs simultaneous multilingual translation and monolingual sequence autoencoding. Positive results from such a strategy were found in (Luong et al., 2016).

Models 20x.19 also perform simultaneous translation and monolingual sequence auto-encoding. However, their architecture is different from 202.16, and they explore different language token prefixing strategies. Compared to 202.16 models, in 20x.19 models, the encoder and decoder layers use 8 attention heads instead of 2. Also, the encoder's input embeddings and the decoder's input and output embeddings are all shared. Finally, the source and target token dictionaries are also shared.

Models 20x.19 explore four approaches of source-side prefix specification (Table 5). As an example, a source segment to be translated from English to Nko is prefixed as follows per model:
206.19: "<to_nqo_Nkoo> "
207.19: "<from_eng_Latn> <to_nqo_Nkoo> "
208.19: "<from> <eng_Latn> <to> <nqo_Nkoo> "
209.19: "<from><eng_Latn><to><nqo_Nkoo> "

4.4 Training

During training, dropout is used with the following probabilities: input token embedding dropout 0.4 (xxx.11) or 0.6 (xxx.16, xxx.19), attention dropout 0.2, ReLU dropout 0.2. The label-smoothed cross-entropy loss function is used with a smoothing rate of 0.2. Optimization is performed using Adam with a weight decay of 0.0001. The inverse squared root learning rate scheduler is used with an initial rate of 1e-7 and 4000 warm-up updates. Gradient clipping is employed with a norm threshold of 1. Effective batches of up to 65,536 tokens are used to train all models. Gradients are accumulated for 1 batch of up 65,536 tokens on A100 GPUs and 4 batches of up to 16384 on Titan XP GPUs before each update.

4.5 Model Selection and Stopping Criteria

Trainings are stopped when BLEU scores on the validation step do not improve after 20K gradient updates. Checkpoints with the highest BLEU scores on the validation set are selected. The average BLEU score across all supported translation directions is used for multilingual model selection.

4.6 Evaluation

CPC/FLoRes-dev and CPC/FLoRes-devtest are respectively used as validation and test sets. For each model, their subsets with languages of interest are considered (see Tables 12 and 11). The chrF++ score, which has been shown to align well with human assessments, especially for morphologically rich languages (Popović, 2017), is used as the main evaluation metric. The Sacre BLEU library (Post, 2018) is used to compute BLEU and chrF++ scores.

4.7 Results

Table 6 shows the test and validation BLEU and chrF++ scores for each model and supported translation direction. The best performing model 208.19 scores 26.00 mean chrF++ on the test set.

Layer Count and Regularization: Compared to 200.11, 200.16 with one extra encoder and decoder layer, and a higher token embedding dropout rate, scored $+0.34$ *nqo* $\leftarrow$ *eng* chrF++.

Multilinguality: Compared to 200.16, the multilingual model 201.16, which has the same architecture and training hyperparameters, scored -0.92 *nqo* $\leftarrow$ *eng* chrF++.

Monolingual Autoencoding: Compared to 201.16, 202.16, which performs simultaneous multilingual translation and monolingual autoencoding, scored $+0.14$ *nqo* $\leftarrow$ *eng* chrF++

Attention Heads and Shared Embeddings: Compared to 202.16, 206.19 which uses 8 attention heads in the encoder and decoder layers, and which shares all input and output embeddings and dictionaries scores $+1.59$ mean chrF++.

Language Token Prefixing: Compared to 206.19, which only specifies target language tokens in the source sequence, 207.19, which specifies the source and target languages as two separate tokens, scored $+0.08$ mean chrF++. 209.19, which specifies the source and target languages as a single

token, scored +0.06 mean chrF++. 208.19, which specifies the source and target languages in a four-token clause scored +0.15 mean chrF++.

5 Discussions

5.1 Fria||el

Improving Usability: As noted by Nko translators, the usability of Fria||el could be improved by: (1) Allowing translators to review their recently submitted tasks before the Workflow Manager proceeds to the next stage of the curation process. (2) Implementing an offline authentication mechanism.

Adding a Translation Memory: Adding a translation memory could increase the productivity, accuracy, and consistency of translators. However, the effect of such a tool on the general quality of translations, including the diversity of synonyms and expression styles should not be overlooked.

Extensibility: Alternate copyediting workflows can be implemented in Fria||el by extending the Workflow Manager and Task Manager. The task presentation user interface can also be adapted to other text curation tasks, such as syntax annotation.

5.2 Parallel Corpora

Handling Short Sequences: The segments in *nicolingua-0005* are, on average, significantly shorter than those in FLoRes and NLLB-SEED. Despite being short, sequences such as ones from the Nko-Français dictionary and Unicode CLDR files, are too valuable to discard. To prevent biasing models towards shorter sequence lengths, we repeated the *(nqo_Nkoo, eng_Latn)* data from CPC/NLLB-SEED five times in the training set. A more principled approach should be considered.

Punctuations, Case and Diacritical Marks: Our models showed sensitivity to minor changes in Latin case, and punctuation as well as Nko diacritical marks (see Appendix G). Including augmented data with lowered case and stripped punctuation and diacritical marks in source sequences in the training corpora may help address this issue.

Learning from Translator Edits: Translator edits, as recorded by Fria||el throughout the copy-edit process, could be useful for various modeling and quality estimation tasks. This data could also be used for an auxiliary copy-edit reconstruction task that may improve the accuracy of a multitask NMT

		Intl. BLEU		chrF++	
model	direction	valid	test	valid	test
200.11	*nqo* ← *eng*	5.40	5.11	28.80	29.73
200.16	*nqo* ← *eng*	5.85	5.25	29.06	30.07
201.16	*nqo* → *bam*	1.19	1.12	16.73	17.04
201.16	*nqo* ← *bam*	2.86	3.19	22.07	23.01
201.16	*nqo* → *eng*	3.65	3.78	26.31	26.99
201.16	*nqo* ← *eng*	6.11	5.71	28.64	29.15
201.16	*nqo* → *fra*	2.33	2.35	22.27	22.61
201.16	*nqo* ← *fra*	4.50	4.29	25.55	25.89
201.16	mean	3.44	3.41	23.60	24.12
202.16	*nqo* → *bam*	1.14	1.00	16.68	16.82
202.16	*nqo* ← *bam*	2.83	3.11	22.33	23.11
202.16	*nqo* → *eng*	4.27	4.26	26.86	27.61
202.16	*nqo* ← *eng*	6.18	5.80	28.63	29.44
202.16	*nqo* → *fra*	2.31	2.74	22.46	22.89
202.16	*nqo* ← *fra*	4.18	4.51	25.22	25.68
202.16	mean	3.49	3.57	23.70	24.26
206.19	*nqo* → *bam*	1.69	1.50	19.04	19.34
206.19	*nqo* ← *bam*	3.63	3.43	23.26	23.81
206.19	*nqo* → *eng*	5.22	5.15	28.70	28.85
206.19	*nqo* ← *eng*	6.79	6.50	29.97	30.66
206.19	*nqo* → *fra*	3.28	3.41	25.26	25.42
206.19	*nqo* ← *fra*	4.77	5.05	26.59	27.03
206.19	mean	4.23	4.17	25.47	25.85
207.19	*nqo* → *bam*	1.52	1.53	19.09	**19.43**
207.19	*nqo* ← *bam*	3.56	3.28	23.10	23.77
207.19	*nqo* → *eng*	5.10	5.05	28.61	28.69
207.19	*nqo* ← *eng*	6.96	6.21	29.92	30.45
207.19	*nqo* → *fra*	3.26	3.51	25.44	**25.98**
207.19	*nqo* ← *fra*	5.23	5.14	26.89	**27.25**
207.19	mean	4.27	4.12	25.51	25.93
208.19	*nqo* → *bam*	1.44	1.52	18.83	19.08
208.19	*nqo* ← *bam*	3.32	3.37	23.38	**24.00**
208.19	*nqo* → *eng*	4.78	5.05	28.64	**29.13**
208.19	*nqo* ← *eng*	6.99	6.44	30.05	**30.83**
208.19	*nqo* → *fra*	3.20	3.61	25.15	25.79
208.19	*nqo* ← *fra*	5.04	4.78	26.73	27.17
208.19	mean	4.13	4.13	25.46	**26.00**
209.19	*nqo* → *bam*	1.60	1.47	19.00	19.25
209.19	*nqo* ← *bam*	3.45	3.43	23.29	23.80
209.19	*nqo* → *eng*	5.07	4.79	28.67	28.82
209.19	*nqo* ← *eng*	6.96	6.58	30.10	30.78
209.19	*nqo* → *fra*	3.49	3.13	25.39	25.76
209.19	*nqo* ← *fra*	5.13	4.92	26.56	27.06
209.19	mean	4.28	4.05	25.50	25.91

Table 6: Our bilingual and multilingual models measured for accuracy on FLoRes-dev (valid) and FLoRes-devtest (test) using the Intl. BLEU (Sacre BLEU with Unicode-aware tokenization) and chrF++ metrics.

model. Finally, translator edit data can be used to train and align translators on consistency standards.

5.3 Neural Machine Translation

Tokenization: Subword regularization, as discussed in (Kudo, 2018) and the dropout-based approach presented by (Provilkov et al., 2020), may lead to increased translation performance for Nko.

Language Token Prefixes: The choice of source-side prefixing strategy had a marginal impact on translation accuracy. Our best model employs a four-token prefix, consisting of source and target language tokens joined with the '<from>' and '<to>' tokens. Our results and those of (Wicks and Duh, 2022), suggest that the specification of translation directions as source-side prefixes in multilingual NMT models merits further investigation.

Learning from Monolingual Data: The use of monolingual Nko data in 202.16 led to marginal improvements in most translation directions. Additional unsupervised tasks such as masked language modeling and denoising should also be explored.

Data Augmentation: Back-translation-based data augmentation, and the generalized data augmentation method in (Xia et al., 2019) could significantly increase NMT performance for Nko.

International BLEU Our BLEU scores are computed with sacreBLEU using international tokenization because sacreBLEU's current default tokenizer (v13a) is inappropriate for Nko; it doesn't properly interpret the Nko Unicode block, particularly its punctuations, to detect word boundaries.

BLEU vs chrF++ The BLEU scores of our models are rather low. This was surprising given the training data size and given Nko translators' feedback on generated translations. This observation is in line with (Popović, 2017)'s hypothesis that chrF++ correlates better with human judgment than BLEU for morphologically rich languages.

6 Conclusion

This work presented Fria||el, a collaborative parallel text curation system with copyediting-based quality workflows. Fria||el enabled the extension of existing multilingual corpora, FLoRes-200 and NLLB-SEED with high-quality Nko translations. Those, and a new corpus we introduced, *nicolingua-0005*, served to build baseline bilingual and multilingual NMT systems for Nko, with the best model achieving the accuracy of 30.84 *eng_Latn* → *nqo_Nkoo* chrF++. We have released Fria||el to facilitate the development and extension of multilingual parallel corpora to more languages. We have also released resources and tools to enable the reproducibility of our results, and further progress towards usable MT systems for Nko.

Acknowledgements

Sources of support that made this project possible include: A generous unrestricted research gift from Meta Platforms, Inc., the Stanford Graduate Fellowship (SGF), the Stanford Natural Language Processing (NLP) group, FriaSoft, and Nko USA Inc. We extend our deep gratitude to FriaSoft for selflessly donating numerous software engineering hours and to Nko USA Inc. for providing Nko language expertise and resources. The commitment of Nko USA Inc. and FriaSoft to advancing West African language technology was instrumental in realizing this project. We thanks to John Hewitt and Tolúlọpẹ́ Ògúnrèmí for giving invaluable feedback. We are grateful to Asmaou Diallo, Nantenin Camara, Koulako Camara, Moussa Thomas Doumbouya, and Ibrahima Doumbouya for creating a supportive environment conducive to the execution of portions of this work in Conakry, Fria, and Boké.

References

Kelly Duke Bryant. 2020. Education and Politics in Colonial French West Africa. In *Oxford Research Encyclopedia of African History*.

Sebastian Burckhardt et al. 2014. Principles of Eventual Consistency. *Foundations and Trends® in Programming Languages*, 1(1-2):1–150.

Laye Camara. 1953. *L'enfant Noir*. Éditions Plon.

Nafadji Sory Conde. 2017. *Introduction au N'ko: Une Alternative Linguistique pour l'Afrique*. Presses de l'Université Kofi Annan and Harmattan Guinée.

Marta R Costa-jussà, Roger Creus, Oriol Domingo, Albert Domínguez, Miquel Escobar, Cayetana López, Marina Garcia, and Margarita Geleta. 2020. Mt-Adapted Datasheets for Datasets: Template and Repository. *arXiv preprint arXiv:2005.13156*.

Baba Mamadi Diane. 2022. Kanjamadi – Kanjamadi for Nko. `https://web.archive.org/web/20231011145800/https://kanjamadi.org/baju/`. Accessed on 2023-10-11.

Diane, Baba Mamadi. 2021. Translation of the Meanings of the Noble Qur'an - N'ko

Translation. https://quranenc.com/en/browse/ankobambara_dayyan/1. Accessed on 2023-10-23.

Coleman Donaldson. 2017. *Clear Language: Script, Register and the N'ko Movement of Manding-Speaking West Africa*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA. Archived from the original on 2019-02-21. Retrieved 2019-02-21.

Coleman Donaldson. 2019. Linguistic and Civic Refinement in the N'ko Movement of Manding-Speaking West Africa. *Signs and Society*, 7:156 – 185.

Moussa Koulako Bala Doumbouya. 2022. Detransliterator. https://github.com/mdoumbouya/detransliterator.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2023. Ethnologue: Languages of the world. Online version.

Christian Federmann. 2018. Appraise Evaluation Framework for Machine Translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128–news Test References for MT Evaluation of 128 Languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24.

Firebase Documentation Google. 2023. Access data offline. https://firebase.google.com/docs/firestore/manage-data/enable-offline. Accessed on 2023-08-04.

International Center, Noor. 2018. Translation of the Meanings of the Noble Qur'an - French Translation. https://quranenc.com/en/browse/french_montada/1. Accessed on 2023-08-15.

International, Saheeh. 2022. Translation of the Meanings of the Noble Qur'an - English Translation. https://quranenc.com/en/browse/english_saheeh/1. Accessed on 2023-08-15.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Paul Amon Kotey. 1975. The Official Language Controversy: Indigenous versus Colonial.

Taku Kudo. 2018. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task Sequence to Sequence Learning. In *International Conference on Learning Representations*.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios Gonzales, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, et al. 2021. Findings of the Americasnlp 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217.

MDN Contributors Mozilla. 2023a. CacheStorage - Web APIs | MDN. https://developer.mozilla.org/en-US/docs/Web/API/CacheStorage. Accessed on 2023-08-04.

MDN Contributors Mozilla. 2023b. Service Worker API - MDN Web Docs. https://developer.mozilla.org/en-US/docs/Web/API/Service_Worker_API. Accessed on 2023-08-04.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory Research for Low-Resourced Machine Translation: A Case Study in African Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Djibril Tamsir Niane. 1974. Histoire et Tradition Historique du Manding. *Présence africaine*, (1):59–74.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp

Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left behind: Scaling Human-Centered Machine Translation.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Dianne Oyler. 2002. Re-Inventing Oral Tradition: The Modern Epic of Souleymane Kanté. *Research in African Literatures*, 33(1):75–93.

Narendra Babu Pamula, K Jairam, and B Rajesh. 2014. Cache-aside Approach for Cloud Design Pattern. *International Journal of Computer Science and Information Technologies*, 5(2):1423–1426.

Maja Popović. 2017. chrf++: Words Helping Character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-dropout: Simple and Effective Subword Regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online. Association for Computational Linguistics.

RFI. 2016. Mandenkan, la Vitalité d'une Langue. https://www.rfi.fr/fr/afrique/20161018-mandenkan-vitalite-une-langue. Accessed on 2023-10-11.

Tina Rosenberg. 2011. Everyone Speaks Text Message - The New York Times. https://www.nytimes.com/2011/12/11/magazine/everyone-speaks-text-message.html. Accessed on 2023-10-23].

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

SIL International & United Bible Societies. 2023. Paratext. https://paratext.org/. Accessed on 2023-10-17.

Yew Chong Soh, Ximena V. Del Carpio, and Liang Choon Wang. 2021. *The Impact of Language of Instruction in Schools on Student Achievement: Evidence from Malaysia using the Synthetic Control Method*. Policy Research Working Papers. The World Bank.

Unicode. 2023a. Unicode CLDR Project. https://cldr.unicode.org/. Accessed on 2023-06-15.

Unicode. 2023b. Unicode cldr project - acknowledgments. https://cldr.unicode.org/index/acknowledgments. Accessed on 2023-06-15.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. *Advances in neural information processing systems*, 30.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Valentin Vydrin, Andrij Rovenchak, and Kirill Maslinsky. 2016. Maninka Reference Corpus: A Presentation. In *TALAf 2016: Traitement Automatique des Langues Africaines (Écrit et Parole). Atelier JEP-TALN-RECITAL 2016-Paris le*.

w3.org. 2022. Service Workers. https://www.w3.org/TR/service-workers/. Accessed on 2023-08-04.

Rachel Wicks and Kevin Duh. 2022. The Effects of Language Token Prefixing for Multilingual Machine Translation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 148–153, Online only. Association for Computational Linguistics.

Wikimedia. 2023. List of Wikipedias. https://meta.wikimedia.org/wiki/List_of_Wikipedias. Accessed on 2023-10-14.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized Data Augmentation for Low-Resource Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, volume 57.

Appendices

A Fria||el User Study Feedback Questionnaire

The feedback questionnaire sent to N'Ko translators appears in Figure 6.

Parallel Data Curation Software Feedback

1. Usability of the software
    a. Is the software useful? Why?
    b. What are your favorite features of the software?
    c. What are some improvements that would make the software better?
2. Translation
    a. Did the software make the translation effort easier? How?
    b. How does the software compare to previous systems you used for translation?
    c. What are some difficulties that you encountered when performing translation tasks?
    d. What are some improvements that would make the software better for translation?
3. Verification
    a. Was the software helpful for performing verification tasks? How?
    b. How does the software compare to previous systems you used for verification?
    c. What are some difficulties that you encountered when performing translation tasks?
    d. What are some improvements that would make the software better for verification?
4. Mistranslations
    a. When performing verifications, what are some frequent types of translation mistakes you found?
    b. Why were these types of mistakes frequent?
    c. Did you communicate with other translators about those types of mistakes?
5. Disagreements
    a. Were there any disagreements regarding language standards?
    b. How were those disagreements resolved?

Figure 6: Survey questions sent to translators after they translated flores-200, nllb-seed, and ntrex-128 to N'Ko

B Fria||el Software Engineering Diagrams

On the next three pages appear:

- Workflow and task management sequence diagrams
- Workflow and Task State-Transition Diagrams
- Logical Data Model
- Physical Data Storage Model in Google Firestore

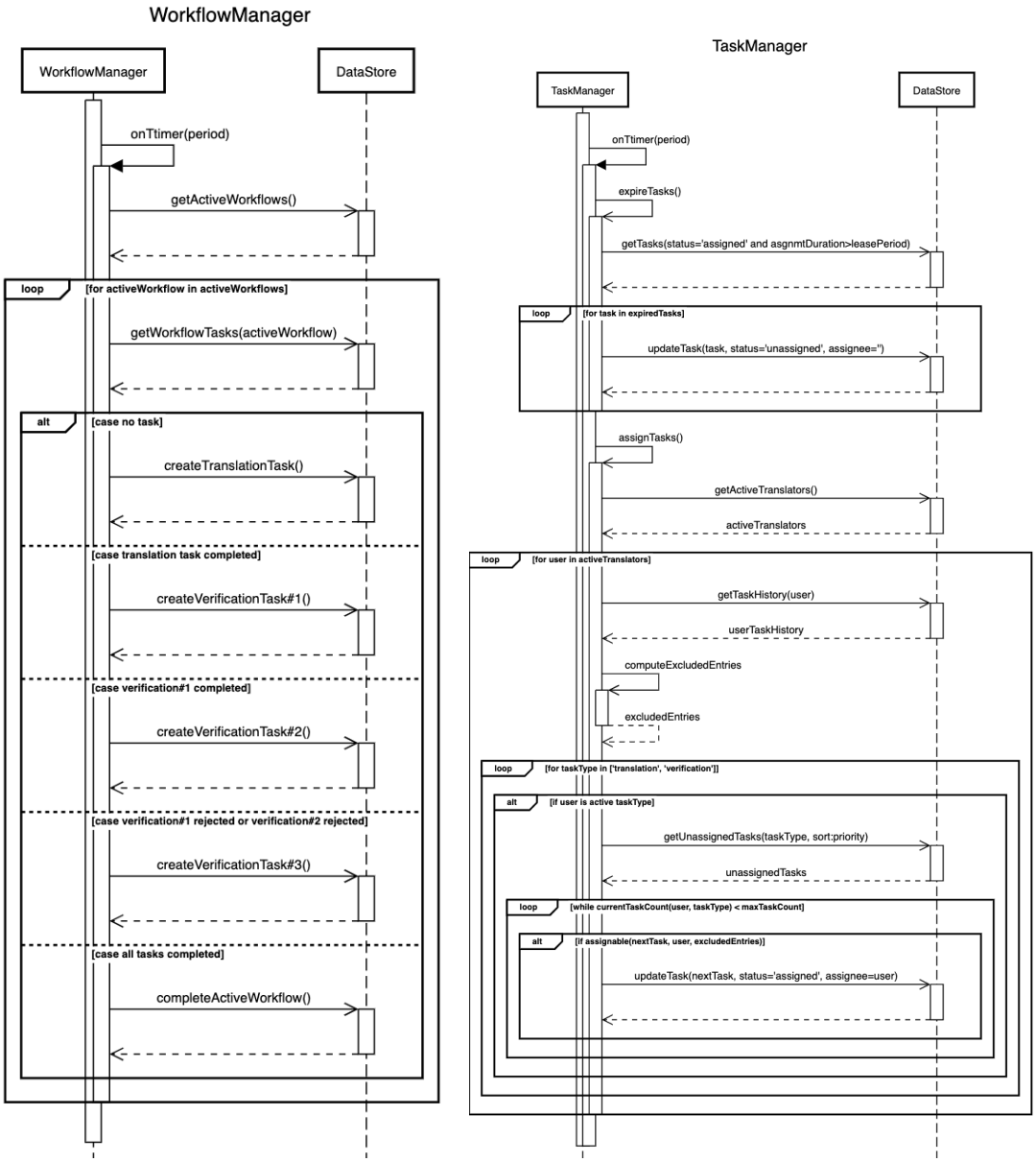


Figure 7: Sequence Diagram: Workflow Manager and Task Manager

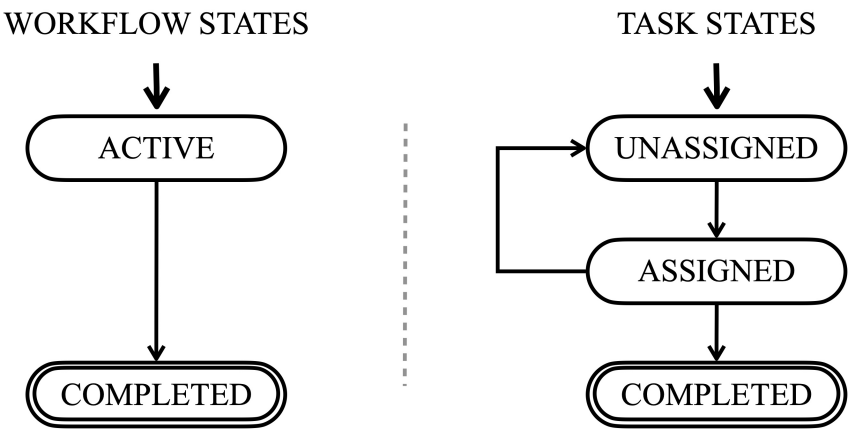


Figure 8: State Transition Diagrams for Tasks and Workflows. A translation workflow entity in the *active* state is created for each dataset entry. The workflow manager creates related *unassigned* tasks as needed, per the rules of the workflow. The Task Manager assigns tasks to users as appropriate. Uncompleted tasks are moved back to the *unassigned* status when not completed within the lease period. The workflow manager moves workflows to the *completed* status when all related tasks are *completed* and there is no need to create additional tasks.

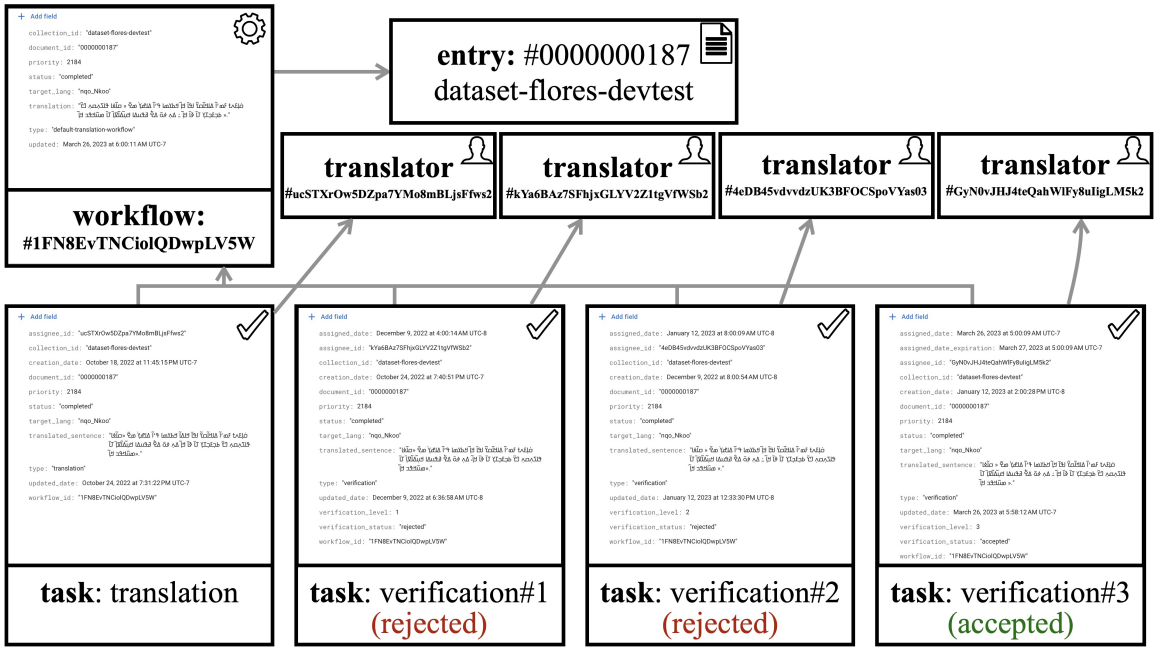


Figure 9: Logical model of entities involved in the curation process of entry#187 of the FLoRes-devtest dataset. Each entity is stored as a document in the Firestore database. The Workflow Manager created one translation task, and three verification tasks, each assigned to a different translator. The third verification task was created because at least one of the previous two resulted in translator edits. Arrows point from referencing to referenced documents.

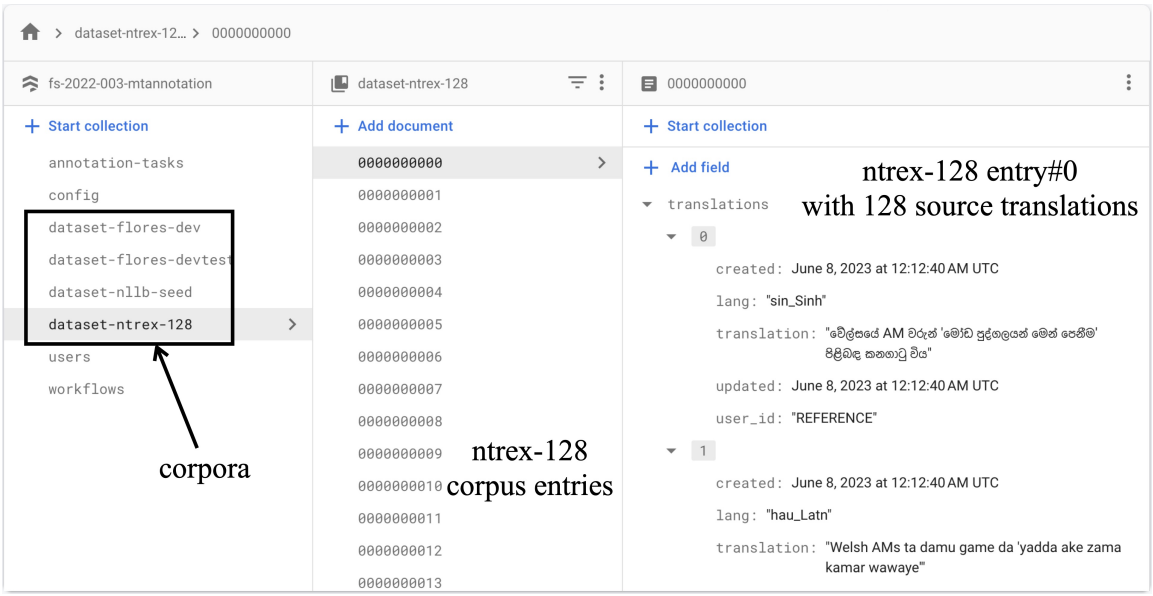


Figure 10: Data Storage in Google Firestore. Each corpus is stored as a collection of documents (left), each of which is identified by its position in the original data files (middle). Each entry contains an array of source translations. Each translation is labeled with its language and script codes (ISO-639_ISO-15924) (right). The system also uses the *users*, *config*, *workflows* and *annotation-tasks* for user, configurations, and data curation workflow management.

C NLLB-SEED *bam_Latn* Quality Issues

Examples of quality issues in NLLB-SEED *bam_Latn* data file appear on the following page.

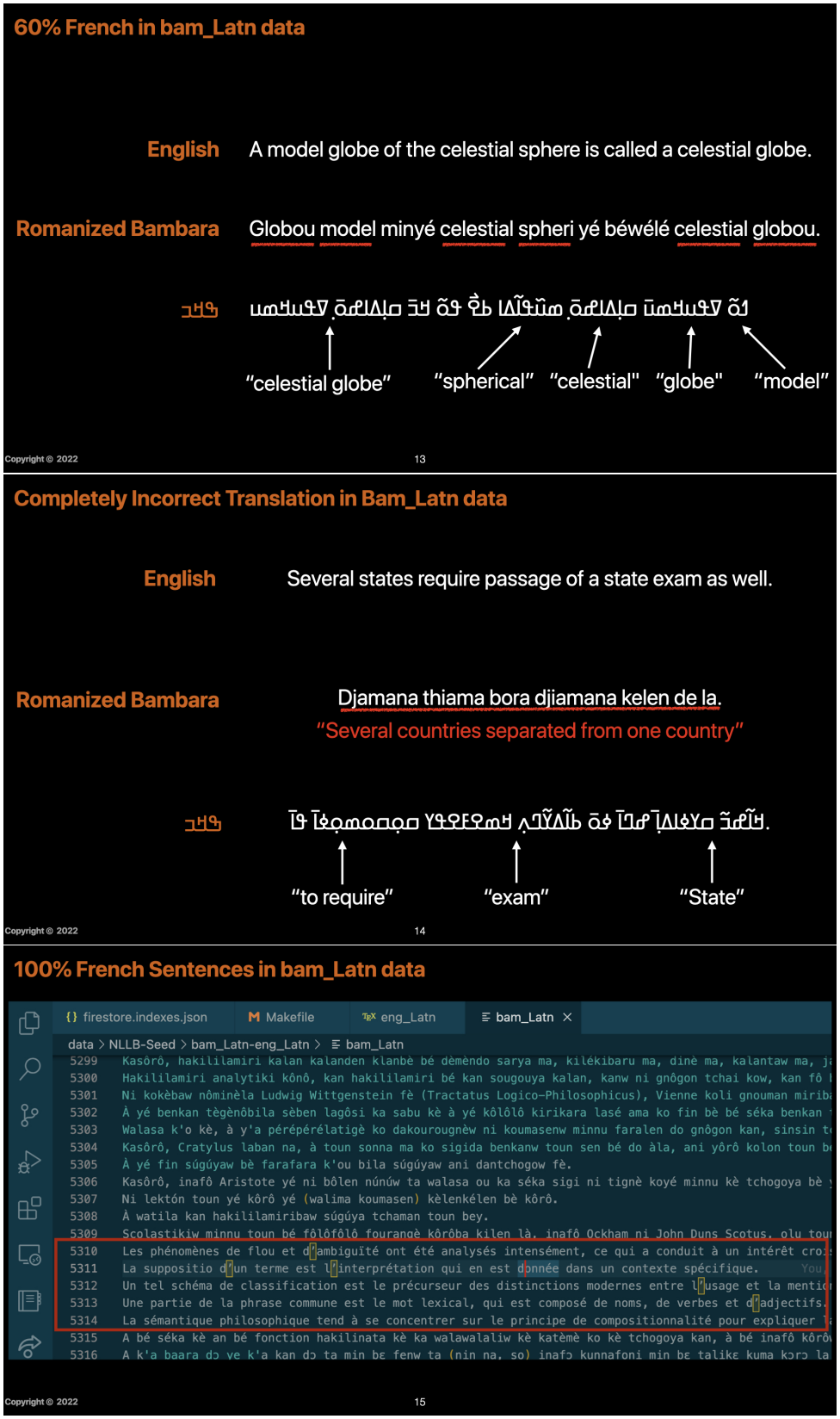


Figure 11: Examples of quality issues in NLLB-SEED *bam_Latn* data file. (1) Sentence with 60% borrowed French words. (2) Incorrect translation. (3) a block of sentences entirely in French. Notice that tonal marks are missing from *bam_Latn* text.

D nicolingua-0005 details

This section provides details on the monolingual, bilingual and trilingual parallel corpora, donated by Nko community members, collectively making up the nicolingua-0005 corpus.

D.0.1 Trilingual Corpora
$(nqo\_Nkoo, eng\_Latn, fra\_Latn)$

baba_mamadi_diane_parallel_002 This corpus is composed of parallel Quran translations in Nko (Diane, Baba Mamadi, 2021), English (International, Saheeh, 2022), and French (International Center, Noor, 2018). The Quran's translation in Nko was originally performed by Baba Mamadi Diane for Islamic education purposes.

kalo_mory_diane_parallel_00{1,2,3} This corpus contains various short phrases collected and translated by Kalo Mory Diane for the purpose of machine translation system development.

solo_farabado_cisse_parallel_002 This corpus contains various short phrases collected and translated by Solo Farabado Cisse for the purpose of machine translation system development.

solo_farabado_cisse_parallel_001 Nko localization strings from the Unicode Common Locale Data Repository (CLDR) (Unicode, 2023a) to which Solo Farabado Cisse and Baba Mamadi Diane contributed (Unicode, 2023b). Corresponding CLDR strings in Nko, English, and French were compiled to make this trilingual parallel corpus.

D.0.2 Bilingual Corpora
(nqo_Nkoo, eng_Latn)

baba_mamadi_diane_parallel_003 This corpus contains segments manually chunked from the Quran and translated by Baba Mamadi Diane specifically for the purpose of creating a corpus usable for machine translation system development.

baba_mamadi_diane_parallel_004 This corpus contains the localization strings of a custom Android build translated by Baba Mamadi Diane.

djibrila_diane_parallel_003 This corpus contains short phrases collected and translated by Djibrila Diane. The phrases also include some basic scientific terminology. The corpus was originally created for education purposes only.

djibrila_diane_parallel_001 This corpus contains short phrases in various tenses collected and translated by Djibrila Diane to serve of MT system development.

djibrila_diane_parallel_002 This corpus contains various short phrases composed and translated by Djibrila Diane for the purpose of MT system development.

D.0.3 Bilingual Corpora
(nqo_Nkoo, fra_Latn)

baba_mamadi_diane_parallel_001 Nko-French dictionary authored by Baba Mamadi Diane for education purposes. Dictionary entries in french with multiple forms (e.g. gender) were automatically expanded using regular expressions.

nafadji_sory_conde_parallel_001 This corpus contains various short phrases composed and translated by Nafadji Sory Conde for the purpose of machine translation system development.

nafadji_sory_conde_parallel_003 This corpus contains phrases from Camara Laye's 1953 novel "L'enfant Noir" (Camara, 1953). The translation was originally done by Nafadji Sory Conde for the purpose of expanding available literature in Nko.

nafadji_sory_conde_parallel_002 This corpus contains various phrases related to Guinean society and sociology. It was created by Nafadji Sory Conde for the purpose of MT system development.

nafadji_sory_conde_parallel_004 This corpus contains segments extracted from the Guinean constitution. It was originally translated by Nafaji Sory Conde for education purposes.

D.0.4 Monolingual Corpora (nqo_Nkoo)

nafadji_sory_conde_monolingual_001 This corpus, composed by Nafadji Sory Conde and his collaborators, contains extracts of books and newspapers in Nko. A substantial part of the corpus was harvested from Kanjamadi.com. This corpus may overlap with the Maninka Reference Corpus (Vydrin et al., 2016).

baba_mamadi_diane_monolingual_00{1,2} These corpora were extracted from various Nko books and articles in various domains including history, religion, philosophy, literature and Science. The corpora were originally composed by Baba M. Diane for the purpose of auto-completion algorithm development for Nko.

lines	words	file	originator	description
6236	175382	baba_mamadi_diane_parallel_002.nqo_Nkoo		
6236	151323	baba_mamadi_diane_parallel_002.eng_Latn	Baba Mamadi Diane	Traductions of the Quran
6236	171085	baba_mamadi_diane_parallel_002.fra_Latn		
7001	28626	kalo_mory_diane_parallel_001.nqo_Nkoo		
7001	17558	kalo_mory_diane_parallel_001.eng_Latn	Kalo Mory Diane	Short Phrases
7001	21593	kalo_mory_diane_parallel_001.fra_Latn		
4001	18864	kalo_mory_diane_parallel_003.nqo_Nkoo		
4001	12891	kalo_mory_diane_parallel_003.eng_Latn	Kalo Mory Diane	Short Phrases
4001	15050	kalo_mory_diane_parallel_003.fra_Latn		
3999	17903	kalo_mory_diane_parallel_002.nqo_Nkoo		
3999	12237	kalo_mory_diane_parallel_002.eng_Latn	Kalo Mory Diane	Short Phrases
3999	14495	kalo_mory_diane_parallel_002.fra_Latn		
3052	13420	solo_farabado_cisse_parallel_002.nqo_Nkoo		
3052	9615	solo_farabado_cisse_parallel_002.eng_Latn	Solo Farabado Cisse	Short Phrases
3052	11308	solo_farabado_cisse_parallel_002.fra_Latn		
1559	2739	solo_farabado_cisse_parallel_001.nqo_Nkoo		
1559	2382	solo_farabado_cisse_parallel_001.eng_Latn	Solo Farabado Cisse	Unicode CLDR Strings
1559	2338	solo_farabado_cisse_parallel_001.fra_Latn		

Table 7: nicolingua-0005's trilingual subsets in Nko (*nqo_Nkoo*), English (*eng_Latn*) and French (*fra_Latn*)

lines	words	file	originator	description
21590	154238	baba_mamadi_diane_parallel_003.nqo_Nkoo	Baba Mamadi Diane	Segments Chunked from the Quran
21590	133369	baba_mamadi_diane_parallel_003.eng_Latn		
36211	119536	baba_mamadi_diane_parallel_004.nqo_Nkoo	Baba Mamadi Diane	Localization Strings for a Custom Android Build
36211	72612	baba_mamadi_diane_parallel_004.eng_Latn		
492	4666	djibrila_diane_parallel_003.nqo_Nkoo	Djibrila Diane	Various Short Phrases and Basic Sci. Terms
492	4122	djibrila_diane_parallel_003.eng_Latn		
1001	3536	djibrila_diane_parallel_001.nqo_Nkoo	Djibrila Diane	Short Phrases in Various Tenses
1001	3487	djibrila_diane_parallel_001.eng_Latn		
148	1303	djibrila_diane_parallel_002.nqo_Nkoo	Djibrila Diane	Various Short Phrases
148	1361	djibrila_diane_parallel_002.eng_Latn		

Table 8: nicolingua-0005's bilingual subsets in Nko (*nqo_Nkoo*) and English (*eng_Latn*)

lines	words	file	originator	description
37894	40436	baba_mamadi_diane_parallel_001.nqo_Nkoo	Baba Mamadi Diane	Nko-Francais Dictionary
37894	41598	baba_mamadi_diane_parallel_001.fra_Latn		
3604	39020	nafadji_sory_conde_parallel_001.nqo_Nkoo	Nafadji Sory Conde	Various Short Phrases
3604	35037	nafadji_sory_conde_parallel_001.fra_Latn		
1141	22379	nafadji_sory_conde_parallel_003.nqo_Nkoo	Nafadji Sory Conde	Segment from "L'enfant Noir"
1141	21049	nafadji_sory_conde_parallel_003.fra_Latn		
2200	16091	nafadji_sory_conde_parallel_002.nqo_Nkoo	Nafadji Sory Conde	Phrases related to Guinean Society and Sociology
2200	15413	nafadji_sory_conde_parallel_002.fra_Latn		
721	11863	nafadji_sory_conde_parallel_004.nqo_Nkoo	Nafadji Sory Conde	Guinean Constitution
721	11345	nafadji_sory_conde_parallel_004.fra_Latn		

Table 9: nicolingua-0005's bilingual subsets in Nko (*nqo_Nkoo*) and French (*fra_Latn*)

lines	words	file	originator	description
134000	2017158	nafadji_sory_conde_monolingual_001.nqo_Nkoo	Nafadji Sory Conde	Various Books and News Papers
44604	853464	baba_mamadi_diane_monolingual_002.nqo_Nkoo	Baba Mamadi Diane	Various Books and Articles
10195	420749	baba_mamadi_diane_monolingual_001.nqo_Nkoo	Baba Mamadi Diane	Various Books and Articles

Table 10: nicolingua-0005's monolingual subsets in Nko (*nqo_Nkoo*)

E Datasheet Questionnaire for *nicolingua-0005*

E.1 Motivation

E.1.1 Who created the dataset(e.g., which team, research group) and on behalf of which entity (e.g. company, institution, organization)?

nicolingua-0005 was curated by Moussa Doumbouya (Stanford University). Its constituent corpora were provided by the following members of Nko USA Inc: Baba Mamadi Diane, Solo Farabado Cisse, Djibrila Diane, Nafadji Sory Conde, Kalo Mory Diane.

E.1.2 Did they fund it themselves? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Nko community members voluntarily composed the included corpora.

E.1.3 For what purpose was the data set created? Was there a specific task in mind? If so, please specify the result type (e.g. unit) to be expected.

Some included corpora were composed specifically for the development of MT systems while others were originally created for educational purposes. See Appendix D for details.

E.1.4 Could any of these uses, or their results, interfere with human will or communicate a false reality?

Not to the best of our knowledge.

E.1.5 What is the antiquity of the file? Provide, please, the current date.

July 19 2023.

E.1.6 Has there been any monetary profit from the creation of this dataset?

No.

E.2 Composition

E.2.1 Is there any synthetic data in the dataset? If so, in what percentage?

The corpus doesn't contain any synthetic data.

E.2.2 Are there multiple types of instances or is there just one type? Please specify the type(s), e.g. Raw data, preprocessed, symbolic.

The corpus contains monolingual and parallel text corpora.

E.2.3 What do the instances (of each type, if appropriate) that comprise the data set represent? (e.g. documents, photos, people, countries).

The instances represent segments of text in Nko, English, and French.

E.2.4 How many instances (of each type, if appropriate) are there in total?

See Tables 4, 7, 8, 9 and 10

E.2.5 Does the dataset contain all possible instances or is it just a sample of a larger set? i.e. Is the dataset different than an original one due to the preprocessing process? In case this dataset is a subset of another one, is the original dataset available?

This dataset is a collection of corpora from various sources. Some sources were integrally sampled (e.g. quran), while other sources were composed by individual translators.

E.2.6 Is there a label or a target associated with each of the instances? If so, please provide a description.

The multilingual subsets of the corpora are matching segments of text in multiple languages.

E.2.7 What is the format of the data? e.g. .json, .xml, .csv .

The files are text files encoded in UTF-8 that have the following extensions matching the iso standard code of the language and writing system they contain: .nqo_Nkoo, .eng_Latn, .fra_Latn.

E.2.8 Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g. because it was unavailable). This does not include intentionally removed information, but might include, e.g. redacted text.

There is no missing information to report.

E.2.9 Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description. Do not include missing information here.

The sentences were benevolently translated by various individuals. A minimal quality control process was adopted during the curation phase. The data may contain some errors. The corpus baba_mamadi_diane_parallel_003 was created by sampling Quran phrases from baba_mamadi_diane_parallel_002. Some parallel Nko segments may be repeated in the monolingual Nko corpora.

E.2.10 Is there any verification that guarantees there is not institutionalization of unfair biases? Both regarding the dataset itself and the potential algorithms that could use it.

no.

E.2.11 Are there recommended data splits, e.g. training, development/validation, testing? If so, please provide a description of these splits explaining the rationale behind them.

The corpora are intended to be used to train natural language processing algorithms.

E.2.12 Is the dataset self-contained, or does it link to or otherwise rely on external resources? e.g., websites, tweets, other datasets. If it links to or relies on external resources, a) Are there any guarantees that they will exist, and remain constant over time? b) Are there official archival versions of the complete dataset? i.e. including the external resources as they existed at the time the dataset was created. c) Are there any restrictions (e.g. licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, if appropriate.

nicolingua-0005 is self-contained.

E.2.13 Does the dataset contain data that might be considered confidential? e.g. data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications. If so, please provide a description.

Not to the best of our knowledge.

E.2.14 Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

Not to the best of our knowledge. Notes: (1) *nicolingua-0005* contains religious text that some people may find offensive or threatening. (2) Some words contained in *nicolingua-0005*, such as the name of certain human body parts included in the Nko-Francais dictionary, may be considered vulgar or offensive.

E.2.15 Does the dataset relate to people? If so, please specify a) Whether the dataset identifies subpopulations or not. b) Whether the dataset identifies indivual people or not. c) Whether it contains information that could vulnerate any individuals or their rights. c) Any other verified information on the topic that can be provided.

The data includes news articles that may reference specific people and people groups. The data also includes literature relating to West African people and people groups and their history.

E.2.16 Does the dataset cover included languages equally?

No. The sizes of various parallel and monolingual subsets have been specified in Table 4.

E.2.17 Is there any evidence that the data may be somehow biased? i.e. towards gender, ethics, beliefs.

The data includes religious texts, articles, and books that may reflect various types of biases. The data may contain biases inherent in historical and current Manding culture such as work organization between men and women, young and old people. Nko doesn't have masculine vs. feminine noun classes. Therefore genders are not distinguished in

Nko nouns and pronouns, which may reduce the potential for gender-based bias.

E.2.18 Is the data made up of formal text, informal text or both equitably?

The data mostly contains formal text.

E.2.19 Does the data contain incorrect language expressions on purpose? Does it contain slang terms? If that's the case, please provide which instances of the data correspond to these.

Not to the best of our knowledge. The dataset may contain unintentional errors.

E.3 Collection Process

E.3.1 Where was the data collected at? Please include as much detail; i.e. country, city, community, entity and so on.

Most data was collected in Conakry, Guinea, and Banakoro, Guinea. Some contributors also worked in Bamako, Mali (Solo F Cisse, Baba M Diane), Egypt (Baba M Diane) and USA (Djibrila Diane) while collecting the datasets.

E.3.2 If the dataset is a sample from a larger set, what was the sampling strategy? i.e. deterministic, probabilistic with specific sampling probabilities.

N/A

E.3.3 Are there any guarantees that the acquisition of the data did not violate any law or anyone's rights?

Not to the best of our knowledge.

E.3.4 Are there any guarantees that prove the data is reliable?

No.

E.3.5 Did the collection process involve the participation of individual people? If so, please report any information available regarding the following questions: Was the data collected from people directly? Did all the involved parts give their explicit consent? Is there any mechanism available to revoke this consent in the future, if desired?

The dataset authors are authors of this paper. They gave their explicit consent.

E.3.6 Has an analysis of the potential impact of the dataset and its use on data subjects been conducted? i.e. a data protection impact analysis. If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

No.

E.3.7 Were any ethical review processes conducted?

No.

E.3.8 Does the data come from a single source or is it the result of a combination of data coming from different sources? In any case, please provide references.

The data was curated from a combination of different sources.

E.3.9 If the same content was to be collected from a different source, would it be similar?

Not Applicable.

E.3.10 Please specify any other information regarding the collection process. i.e. Who collected the data, whether they were compensated or not, what mechanisms were used. Please, only include if verified.

E.4 Preprocessing/Cleaning/Labelling

E.4.1 Please specify any information regarding the preprocessing that you may know (e.g. the person who created the dataset has somehow explained it) or be able to find (e.g. there exists and informational site). Please, only include if verified. i.e. Was there any mechanism applied to obtain a neutral language? Were all instances preprocessed the same way?

The data was normalized with Unicode normalization form NFC: Canonical Decomposition followed by Canonical Composition. Non-Nko characters were stripped from monolingual Nko text. Extra punctuations were removed from some sources. Some entries in Baba Mamadi Diane's Nko-Francais dictionary were expanded using regular expressions so that separate forms of the same

words (e.g. gendered, plural) were repeated as separate entries.

E.5 Uses

E.5.1 Has the dataset been used already? If so, please provide a description.

The data was used to build baseline neural machine translation algorithms for Nko. See Section 4.

E.5.2 Is there a repository that links to any or all papers or systems that use this dataset? If so, please provide a link or any other access point.

`https://github.com/mdoumbouya/nicolingua-0005-nqo-nmt-resources`
`https://github.com/mdoumbouya/nicolingua-0005-nqo-nmt-resources`

E.5.3 What (other) tasks could the dataset be used for? Please include your own intentions, if any.

Any natural language processing tasks including language modeling and machine translation.

E.5.4 Are there tasks for which the dataset should not be used? If so, please provide a description.

Not to the best of our knowledge.

E.6 Distribution

E.6.1 Please specify the source where you got the dataset from.

The datasets came from the following individuals:

E.6.2 When was the dataset first released?

July 19 2023.

E.6.3 Are there any restrictions regarding the distribution and/or usage of this data in any particular geographic regions?

No.

E.6.4 Is the dataset distributed under a copyright or other intellectual property (IP) license? And/or under applicable terms of use (ToU)? Please cite a verified source.

The dataset is openly available under the Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license.

E.7 Maintenance

E.7.1 Is there any verified manner of contacting the creator of the dataset?

The authors of this paper can be contacted via email.

E.7.2 Specify any limitations there might be to contributing to the dataset. i.e. Can anyone contribute to it? Can someone do it at all?

The dataset is openly available under the Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license.

E.7.3 Has any erratum been notified?

No.

E.7.4 Is there any verified information on whether the dataset will be updated in any form in the future? Is someone in charge of checking if any of the data has become irrelevant throughout time? If so, will it be removed or labeled somehow?

The dataset will be maintained on GitHub. Any updates will be made available in the same GitHub repository.

E.7.5 Is there any available log about the changes performed previously in the dataset?

Any future modifications will be tracked in GitHub's version control.

E.7.6 Could changes to current legislation end the right-of-use of the dataset?

Not to the best of our knowledge.

E.7.7 Are there any lifelong learning updates, such as vocabulary enrichment, automatically developed?

No.

F Train, Valid and Test Subset Details

Details on the training, validation, and test subset composition for each model appear on the following page.

TRAIN

lines	words	file	200	201	202-9
6193	148442	common-parallel-corpora/multitext-nllb-seed/bam_Latn		✓	✓
6193	136157	cpc/multitext-nllb-seed/eng_Latn	✓	✓	✓
6193	184138	cpc/multitext-nllb-seed/nqo_Nkoo	✓	✓	✓
6236	151323	nicolingua-0005/baba_mamadi_diane_parallel_002.eng_Latn	✓	✓	✓
6236	171085	nicolingua-0005/baba_mamadi_diane_parallel_002.fra_Latn		✓	✓
6236	175382	nicolingua-0005/baba_mamadi_diane_parallel_002.nqo_Nkoo	✓	✓	✓
7001	17558	nicolingua-0005/kalo_mory_diane_parallel_001.eng_Latn	✓	✓	✓
7001	21593	nicolingua-0005/kalo_mory_diane_parallel_001.fra_Latn		✓	✓
7001	28626	nicolingua-0005/kalo_mory_diane_parallel_001.nqo_Nkoo	✓	✓	✓
4001	12891	nicolingua-0005/kalo_mory_diane_parallel_003.eng_Latn	✓	✓	✓
4001	15050	nicolingua-0005/kalo_mory_diane_parallel_003.fra_Latn		✓	✓
4001	18864	nicolingua-0005/kalo_mory_diane_parallel_003.nqo_Nkoo	✓	✓	✓
3999	12237	nicolingua-0005/kalo_mory_diane_parallel_002.eng_Latn	✓	✓	✓
3999	14495	nicolingua-0005/kalo_mory_diane_parallel_002.fra_Latn		✓	✓
3999	17903	nicolingua-0005/kalo_mory_diane_parallel_002.nqo_Nkoo	✓	✓	✓
3052	9615	nicolingua-0005/solo_farabado_cisse_parallel_002.eng_Latn	✓	✓	✓
3052	11308	nicolingua-0005/solo_farabado_cisse_parallel_002.fra_Latn		✓	✓
3052	13420	nicolingua-0005/solo_farabado_cisse_parallel_002.nqo_Nkoo	✓	✓	✓
1559	2382	nicolingua-0005/solo_farabado_cisse_parallel_001.eng_Latn	✓	✓	✓
1559	2338	nicolingua-0005/solo_farabado_cisse_parallel_001.fra_Latn		✓	✓
1559	2739	nicolingua-0005/solo_farabado_cisse_parallel_001.nqo_Nkoo	✓	✓	✓
21590	133369	nicolingua-0005/baba_mamadi_diane_parallel_003.eng_Latn	✓	✓	✓
21590	154238	nicolingua-0005/baba_mamadi_diane_parallel_003.nqo_Nkoo	✓	✓	✓
36211	72612	nicolingua-0005/baba_mamadi_diane_parallel_004.eng_Latn	✓	✓	✓
36211	119536	nicolingua-0005/baba_mamadi_diane_parallel_004.nqo_Nkoo	✓	✓	✓
1001	3487	nicolingua-0005/djibrila_diane_parallel_001.eng_Latn	✓	✓	✓
1001	3536	nicolingua-0005/djibrila_diane_parallel_001.nqo_Nkoo	✓	✓	✓
148	1361	nicolingua-0005/djibrila_diane_parallel_002.eng_Latn	✓	✓	✓
148	1303	nicolingua-0005/djibrila_diane_parallel_002.nqo_Nkoo	✓	✓	✓
492	4122	nicolingua-0005/djibrila_diane_parallel_003.eng_Latn	✓	✓	✓
492	4666	nicolingua-0005/djibrila_diane_parallel_003.nqo_Nkoo	✓	✓	✓
37894	41598	nicolingua-0005/baba_mamadi_diane_parallel_001.fra_Latn		✓	✓
37894	40436	nicolingua-0005/baba_mamadi_diane_parallel_001.nqo_Nkoo		✓	✓
3604	35037	nicolingua-0005/nafadji_sory_conde_parallel_001.fra_Latn		✓	✓
3604	39020	nicolingua-0005/nafadji_sory_conde_parallel_001.nqo_Nkoo		✓	✓
2200	15413	nicolingua-0005/nafadji_sory_conde_parallel_002.fra_Latn		✓	✓
2200	16091	nicolingua-0005/nafadji_sory_conde_parallel_002.nqo_Nkoo		✓	✓
1141	21049	nicolingua-0005/nafadji_sory_conde_parallel_003.fra_Latn		✓	✓
1141	22379	nicolingua-0005/nafadji_sory_conde_parallel_003.nqo_Nkoo		✓	✓
721	11345	nicolingua-0005/nafadji_sory_conde_parallel_004.fra_Latn		✓	✓
721	11863	nicolingua-0005/nafadji_sory_conde_parallel_004.nqo_Nkoo		✓	✓
134000	2017158	nicolingua-0005/nafadji_sory_conde_monolingual_001.nqo_Nkoo			✓
10195	420749	nicolingua-0005/baba_mamadi_diane_monolingual_001.nqo_Nkoo			✓
44604	853464	nicolingua-0005/baba_mamadi_diane_monolingual_002.nqo_Nkoo			✓

Table 11: Data files included in the training set of each model family

VALID

lines	words	file	200	201	202-9
997	21565	common-parallel-corpora/flores-200-dev/bam_Latn.dev		✓	✓
997	20954	common-parallel-corpora/flores-200-dev/eng_Latn.dev	✓	✓	✓
997	23957	common-parallel-corpora/flores-200-dev/fra_Latn.dev		✓	✓
997	27361	common-parallel-corpora/flores-200-dev/nqo_Nkoo.dev	✓	✓	✓

TEST

lines	words	file	200	201	202-9
1012	22565	common-parallel-corpora/flores-200-devtest/bam_Latn.devtest		✓	✓
1012	21901	common-parallel-corpora/flores-200-devtest/eng_Latn.devtest	✓	✓	✓
1012	25319	common-parallel-corpora/flores-200-devtest/fra_Latn.devtest		✓	✓
1012	29503	common-parallel-corpora/flores-200-devtest/nqo_Nkoo.devtest	✓	✓	✓

Table 12: Data files included in the validation and test sets of each model family

G Examples of Translations

Examples of generations highlighting the sensitivity our ouf baseline NMT system to punctuation and case appear on the following page.

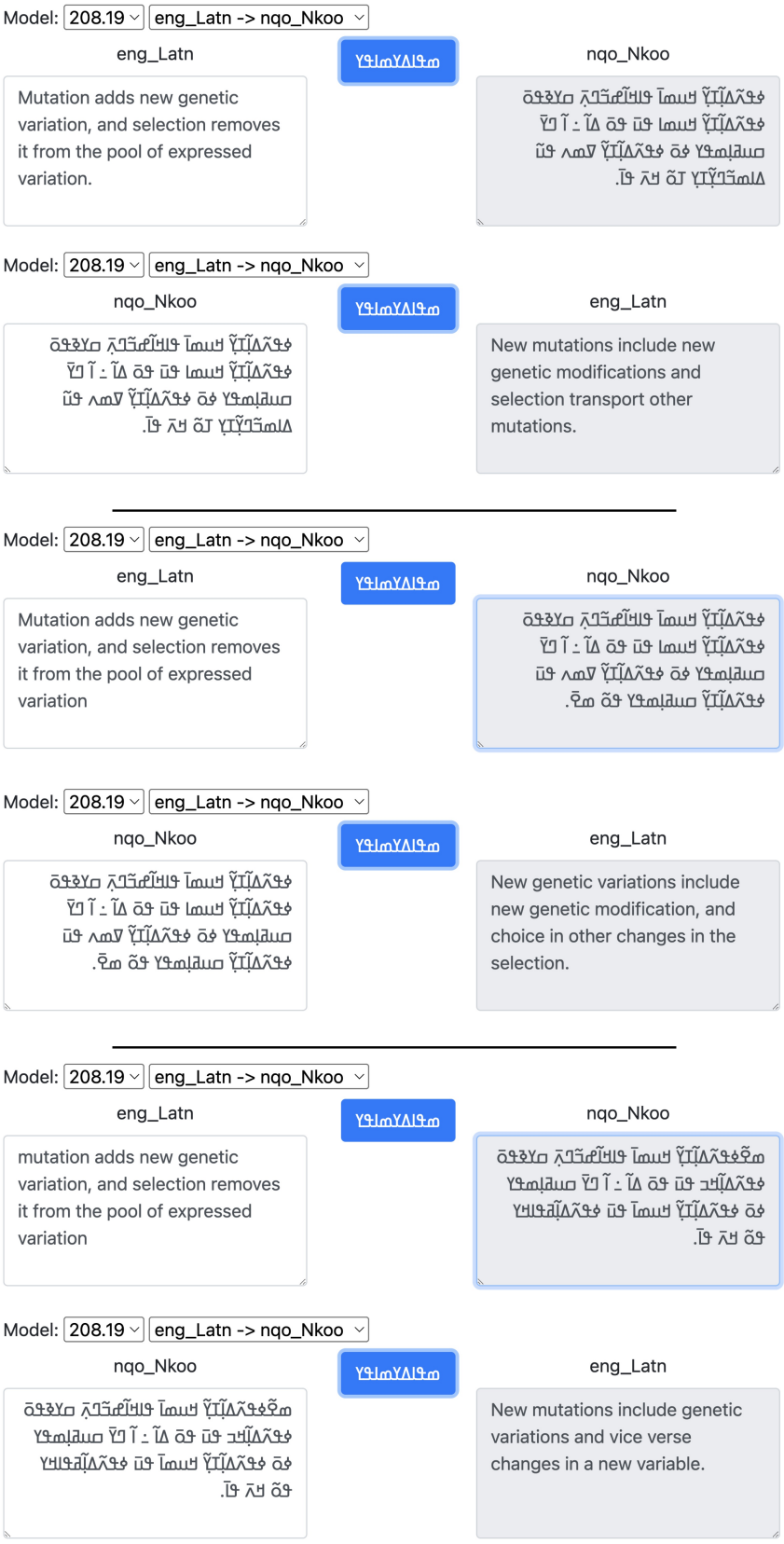


Figure 12: A sentence from the FLoRes-200-devtest corpus translated from English to Nko and back-translated to English using model 208.19. The three examples highlight the sensitivity of our baseline system to punctuation and case. Top: original sentence; Middle: removed final period; Bottom: removed initial capitalization and final period.

TTIC's Submission to WMT-SLT 23

Marcelo Sandoval-Castañeda
TTI-Chicago
marcelo@ttic.edu

Yanhong Li
TTI-Chicago
yanhongli@ttic.edu

Bowen Shi
Meta AI
bshi@meta.com

Diane Brentari
The University of Chicago
dbrentari@uchicago.edu

Karen Livescu
TTI-Chicago
klivescu@ttic.edu

Gregory Shakhnarovich
TTI-Chicago
gregory@ttic.edu

Abstract

We describe TTIC's submission to the WMT 2023 Sign Language Translation shared task on the Swiss-German Sign Language (DSGS) to German track. Our approach explores the advantages of using large-scale self-supervised pre-training in the task of sign language translation, over more traditional approaches that rely heavily on supervision, along with costly labels such as gloss annotations. The proposed model consists of a VideoSwin transformer for image encoding, and a T5 model adapted to receive VideoSwin features as input instead of text. On WMT-SLT 22's development set, this system achieves 2.03 BLEU score, a 59% increase over the previous best reported performance. On the official test set, our primary submission achieves 1.1 BLEU score and 17.0 chrF score. It also achieves the highest human evaluation score among all the participants.

1 Introduction

Sign language translation (SLT) is the task of translating a signed language to a written language, typically the lingua franca of the region the signed language is utilized. In recent years, SLT has received increased attention from the natural language processing (NLP) and computer vision (CV) communities.

The best-performing SLT models primarily rely on glosses (Zhou et al., 2021; Chen et al., 2022), a combination of morpheme translations into the target language along with differentiating phonological features like handshape and location. However, annotating glosses is expensive (Müller et al., 2023b), and recent research has begun to move away from gloss-based translation (Shi et al., 2022a; Uthus et al., 2023; Lin et al., 2023), particularly in regimes where larger datasets are available.

In this paper, we study large-scale self-supervision and noisy supervision for Swiss-German Sign Language (DSGS from the German *Deutschschweizer Gebärdensprache*) to German SLT, as part of the WMT-SLT 23 shared task (Müller et al., 2023a). Given recent findings on self-supervised transformers' perfomance on isolated sign recognition and feature extraction (Sandoval-Castañeda et al., 2023), we utilize a VideoSwin (Liu et al., 2022) visual feature extractor with BEVT pre-training (Wang et al., 2022). Additionally, we use T5 (Raffel et al., 2020) as a sequence-to-sequence translation model into German because of its state-of-the-art performance on American Sign Language (ASL) to English SLT with pose input (Uthus et al., 2023). Depending on the generation algorithm, our model achieves either the highest BLEU score (Papineni et al., 2002) or the highest chrF (Popović, 2015) in the task's leaderboard. With top-k beam sampling, it achieves 0.8 BLEU and 17.3 chrF, and with diverse beam search (Vijayakumar et al., 2016), it achieves 1.1 BLEU and 17.0 chrF.

2 Method

Our model follows the most common gloss-free translation architecture, composed of a visual encoding backbone and a transformer-based model for sequence modeling. Our visual backbone is a Video Swin Transformer (VideoSwin) and our sequence-to-sequence model is a Text-to-Text Transfer Transformer (T5).

2.1 VideoSwin

VideoSwin is an architecture proposed as an extension of the shifted-window transformer (Liu et al., 2021), a hierarchical vision transformer that relies on windowed self-attention for computational efficiency. We pre-train a VideoSwin using video-only BEVT pre-training (Wang et al., 2022) on OpenASL (Shi et al., 2022a), using the codebook from a discrete variational autoencoder (dVAE) (Ramesh et al., 2021) to produce the labels in the self-supervision objective. Though OpenASL is originally a sign language translation dataset, we ig-

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 344–350
December 6–7, 2023.

nore the English translations and train exclusively on the dataset's videos. Then, we fine-tune on the gloss-based version (Dafnis et al., 2022; Neidle and Ballard, 2022) of WLASL2000 (Li et al., 2020) for supervised isolated sign language recognition.

Given a video with dimensions $16 \times 224 \times 224$, that is, 16 frames of height 224 pixels and width 224 pixels, VideoSwin first divides the input into patches of shape $2 \times 4 \times 4$ and produces a 128-dimensional vector representation for each patch, producing a tensor of shape $8 \times 56 \times 56 \times 128$. After the first two windowed self-attention blocks, patch representations are divided into non-overlapping groups of four spatially contiguous patches, which are then projected into a single 256-dimensional vector each. This is done again after two windowed self-attention blocks, and once more after eighteen windowed self-attention blocks. The resulting tensor after these patch merging steps has dimensions $8 \times 7 \times 7 \times 1024$.

For translation, we pad the video at the end such that the number of frames is a multiple of 16, divide it into non-overlapping segments of 16 contiguous frames, and run each segment independently through the model. The visual features extracted from the model are the output of the last windowed self-attention block from VideoSwin for each video segment. Then, we concatenate them across the time dimension, and remove the model's outputs that correspond to the padding frames. This is done both during training and during inference. More formally:

$$f_{1:\lceil T/2 \rceil} = M^v(I_{1:T}) \quad (1)$$

where $I_{1:T}$ is a sequence of T image frames, M^v is our VideoSwin model, and $f_{1:\lceil T/2 \rceil}$ is the resulting sequence of visual features, with dimensions $\lceil T/2 \rceil \times 7 \times 7 \times 1024$.

2.2 T5

T5 is a standard encoder–decoder text transformer (Raffel et al., 2020). Recent research has found that T5 pre-trained on English and fine-tuned for ASL to English translation produces state-of-the-art results using pose input (Uthus et al., 2023). We use a T5 model pre-trained on the German Colossal Cleaned Common Crawl (GC4) corpus, which is a cleaned and pre-processed German-only corpus based on Common Crawl. We take pre-trained checkpoints[1] from HuggingFace (Wolf et al., 2020).

[1] https://huggingface.co/GermanT5

Since our sequence of visual features $f_{1:\lceil T/2 \rceil}$ has dimensions $\lceil T/2 \rceil \times 7 \times 7 \times 1024$, we project these into a single vector per timestep, $\lceil T/2 \rceil \times 1024$. To this end, we use a simple convolutional layer with kernel size $1 \times 7 \times 7$. We replace the word embeddings layer from the T5 model with this convolutional layer. This is the only component trained from scratch in our DSGS to German translation model.

2.3 Training Loss

We use cross-entropy loss for BEVT pre-training, isolated sign language recognition (ISLR) fine-tuning, text-to-text pre-training, and features-to-text translation.

2.4 Inference

We expand on the effect of generation algorithms in Section 4.5. For our primary submission, our generation algorithm of choice is diverse beam search (Vijayakumar et al., 2016), with 5 beams, 5 beam groups, and a diversity penalty of 1.

3 Experimental Setup

3.1 Data

We use both last year's and this year's WMT-SLT datasets. Last year's training dataset is composed of data from FocusNews and SRF, both news TV programs, consisting of 17,207 manually aligned DSGS–German pairs, for a total of 35 hours. German text is obtained from the subtitles that correspond to the original spoken German content, and DSGS video is obtained from live translators. Manual alignment is necessary to ensure that each translated sentence in the video is assigned the correct German sentence. In contrast, this year's dataset consists of 231,834 DSGS–German pairs without any manual alignment, for a total of 437 hours, of only SRF data. Last year's SRF data is a subset of this year's dataset, with the key difference that the superset does not contain manually aligned and verified German translations.

Additionally, we use OpenASL (Shi et al., 2022a), a dataset consisting of 288 hours of ASL-English pairs, for the self-supervised pre-training of our visual encoder. In this pre-training we also employ the labels produced by the codebook of a dVAE, which was separately trained on Conceptual Captions (Sharma et al., 2018). For the second stage of pre-training of our visual encoder,

we fine-tune the pre-trained model on the gloss-based version of WLASL2000 (Li et al., 2020), a 14-hour dataset consisting of 19,673 isolated sign ASL videos and 1535 gloss labels (Neidle and Ballard, 2022).

Lastly, the checkpoint we use for T5 is pre-trained on the GC4 corpus. GC4 is a German-only corpus that contains 40.8 billion tokens in total. This is a subset of Common Crawl where the primary language is German extracted between 2015 and 2021.

3.2 Training

Our visual backbone is VideoSwin's base version. It consists of 88.1 million parameters, and is composed of 2 windowed self-attention blocks with 128 hidden dimensions at stage 1, 2 with 256 hidden dimensions at stage 2, 18 with 512 hidden dimensions at stage 3, and 2 with 1024 dimensions at stage 4. We pre-train it in two stages. First, we train it for 150 epochs on OpenASL via video-only BEVT where the labels are produced by the codebook of a dVAE, with a learning rate of 0.0005 on a cosine schedule with 10 warmup epochs and batch size of 128 across 8 GPUs. We use AdamW (Loshchilov and Hutter, 2019) as the optimizer, with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and 0.05 weight decay. In the second stage, we train it on gloss-based WLASL2000[2] for classification for 120 epochs, this time with a learning rate of 0.0003 on a cosine schedule with 2.5 warmup epochs and a batch size of 256 across 8 GPUs. Again, we use AdamW as our optimizer, with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and 0.001 weight decay. Our VideoSwin backbone is then frozen for the rest of our model's training.

For translation, we adapt T5's efficient-large (Tay et al., 2022) version using a convolutional layer to project our representations. This model is composed of 1.09 billion parameters, with 36 self-attention blocks in the encoder and 36 self-attention blocks in the decoder. To tokenize the target translations, we use a SentencePiece tokenizer trained on the same data as the German-only T5, with a vocabulary size of 32,128. We train it in two stages, using both WMT-SLT 22 and WMT-SLT 23 data. WMT-SLT 23 translations are weakly supervised labels, since there is no guarantee of alignment between the video and the corresponding text translations. Therefore, our pipeline uses it as a large, noisy dataset to train the model which will be eventually further fine-tuned with WMT-SLT 22, which has manually verified labels. First, we train it for 8500 steps on WMT-SLT 23's dataset, with a learning rate of 0.001 on a linearly decreasing schedule and a batch size of 64 across 8 GPUs. We use Adafactor (Shazeer and Stern, 2018) as the optimizer. For the second stage, we train the model for 1500 steps on WMT-SLT 22's dataset, with a learning rate of 0.0002 on a linearly decreasing schedule with a batch size of 64 across 8 GPUs. We also use Adafactor at this stage.

3.3 Evaluation

We evaluate our systems and compare them with last year's submissions, since we use the same validation set, using BLEU-1, BLEU-2, BLEU-3 and BLEU-4.

4 Experimental Results

Table 1 shows the performance of our model on WMT-SLT 22's development set, compared to the highest reported BLEU-4 scores reported on the test set by human evaluation (Müller et al., 2022). We also include MSMUNICH's model based on AV-HuBERT (Shi et al., 2022c), since it achieved the highest BLEU-4 score on the development set. Our model performs at least 81% better than the others in all metrics, and 99% better in BLEU-4, which is the metric used in the challenge's leaderboard.

We additionally perform several ablations to quantify the impact of our model's several moving parts. Our ablations are performed using T5's efficient-base configuration with 619 million parameters for time efficiency, unless otherwise specified.

4.1 Visual Backbone

We first evaluate the effect of our choice of visual backbone and pre-training tasks. We compare our VideoSwin backbone with two other models. First, we take a standard I3D model (Carreira and Zisserman, 2017) trained on the ISLR component of the BBC-Oxford British Sign Language dataset (Albanie et al., 2020), called BSL5K (Varol et al., 2021), since I3D is the most commonly used backbone for SL translation. Previous literature suggests that diversity of isolated signs leads to

[2]The original data can be downloaded here: https://dxli94.github.io/WLASL/ And the gloss-based labels can be downloaded here: https://dai.cs.rutgers.edu/dai/s/aboutwlasl

Model	Backbone	Translation Data	B1	B2	B3	B4
MSMUNICH (Dey et al., 2022)	AV-HuBERT	WMT-SLT 22	–	–	–	1.28
MSMUNICH (Dey et al., 2022)	I3D	WMT-SLT 22	–	–	–	0.77
UZH (Müller et al., 2022)	OpenPose	WMT-SLT 22	–	–	–	0.59
TTIC (Shi et al., 2022b)	I3D	WMT-SLT 22	8.36	2.92	1.55	1.02
Ours	**VideoSwin**	**WMT-SLT 22 + 23**	**15.19**	**5.62**	**3.06**	**2.03**

Table 1: Performance of our model on WMT-SLT 22's development set compared to WMT-SLT 22's highest reported scores. B1, B2, B3, and B4 stand for BLEU-1, BLEU-2, BLEU-3, and BLEU-4, respectively.

better representations for downstream tasks like translation, and BSL5K is the largest and most diverse ISLR dataset to our knowledge. We also include an I3D model trained on WLASL2000 for comparison. Second, we also include a version of our pipeline where we replace OpenASL with the WMT-SLT 23 training data without translations for self-supervised pre-training. However, we do not include WLASL2000 fine-tuning for this model, given the language differences between DSGS and ASL.

As Table 2 shows, There is significant deterioration from shifting our self-supervised BEVT VideoSwin backbone to any of the fully supervised I3Ds. Similarly, despite being pre-trained in a different language, OpenASL pre-training performs much better than WMT-SLT 23 pre-training, despite being the smaller training set (288 vs. 437 hours). This is likely a product of OpenASL's far superior diversity in backgrounds, which are masked in WMT-SLT 23, topics (social media content vs. news), and signers (220 vs. 4).

Backbone	Data	B1	B2	B3	B4
I3D	ASL	12.15	2.96	1.31	0.79
I3D	BSL	12.79	2.80	1.12	0.59
BEVT	DSGS	12.43	3.34	1.72	1.16
BEVT	**ASL**	**15.16**	**5.20**	**2.75**	**1.82**

Table 2: Impact of visual backbone and training data on our model's performance. I3D refers to Inception3D models and BEVT refers to BEVT VideoSwin models. We group our pre-training data by language: BSL refers to BSL5K, DSGS refers to WMT-SLT 23, and ASL refers to OpenASL (if BEVT) and WLASL2000.

4.2 Translation Pre-Training

We also consider different combinations of our two DSGS to German translation datasets. In our training set-up, the model is first trained on WMT-SLT 23's weakly supervised labels, and then fine-tuned on WMT-SLT 22's manually aligned labels. We compare this to settings where we use either only WMT-SLT 23 data or only WMT-SLT 22 data. Using only WMT-SLT 22 data is equivalent to WMT-SLT 22's challenge.

From Table 3, we can see that despite the possible misalignments in WMT-SLT 23, training on a larger set of translation pairs is superior to using only WMT-SLT 22 data. However, the best performance we obtain comes from first training on the potentially noisy but large WMT-SLT 23, and then fine-tuning on WMT-SLT 22 for fewer steps.

W22	W23	B1	B2	B3	B4
✗	✓	14.28	4.33	2.27	1.58
✓	✗	13.47	4.30	2.19	1.42
✓	✓	**15.16**	**5.20**	**2.75**	**1.82**

Table 3: Impact of weak supervision translation labels on our model's performance. W22 refers to training on WMT-SLT 22 data and W23 refers to training on WMT-SLT 23 data. Where both are used, the model is trained on WMT-SLT 23 first and then on WMT-SLT 22.

4.3 Sequence-to-Sequence Model

In addition to T5, we also adapt Whisper (Radford et al., 2023) for DSGS to German translation and test it. The intuition behind it is that audio and video both have a time dimension that corresponds to seconds, whereas text does not. We adapt it in a similar fashion to T5, with the addition of a $4\times$ bicubic interpolation step right before the convolutional layer. We do so because Whisper receives input with 50 tokens per second, whereas our VideoSwin features produce one representation every two frames, for 12.5 every second, since the video is at 25 frames per second.

Results in Table 4 suggest that using a text-to-text model performs significantly better than a speech-to-text one.

Model	B1	B2	B3	B4
Whisper	15.08	4.26	2.04	1.29
T5	**15.16**	**5.20**	**2.75**	**1.82**

Table 4: Impact of sequence-to-sequence component of our model on translation performance.

4.4 Model Size

Next, we consider model size in Table 5. Due to computational and time constraints, we only evaluate T5-efficient-small, T5-efficient-base, and T5-efficient-large, with 142 million, 619 million, and 1.09 billion parameters respectively. As expected, larger models correspond to better performance.

Size	Params	B1	B2	B3	B4
Small	142m	15.43	5.13	2.47	1.52
Base	619m	15.16	5.20	2.75	1.82
Large	**1.09b**	**15.19**	**5.62**	**3.06**	**2.03**

Table 5: Impact of model size on our model's performance.

4.5 Decoding Algorithm

Last, we evaluate the effect of different choices of decoding algorithm on test set performance, using our best performing model, T5-efficient-large. We compare the results generated from the following algorithms: greedy decoding, top-k sampling (Fan et al., 2018), beam search, top-k beam sampling, and diverse beam search (Vijayakumar et al., 2016), with $k = 50$ and beam width set to 5. Table 6 shows our results from this experiment, revealing that diverse beam search and top-k beam sampling represent the most significant improvements from the greedy decoding baseline. We choose diverse beam search for our primary submission to the challenge, as it is the only one that improves both BLEU and chrF scores from our baseline.

Generation Algorithm	B4	chrF
Greedy Decoding	0.9	16.0
Top-k Sampling	0.8	16.3
Beam Search	0.9	17.2
Top-k Beam Sampling	0.8	17.3
Diverse Beam Search	**1.1**	**17.0**

Table 6: Impact of generation algorithm for our best model in WMT-SLT 23's test set.

5 Conclusion

Our experiments evaluate a hierarchical vision transformer on the task of sign language translation for the first time, and demonstrate superior performance over I3D-based translation models. We also show the benefits of using large datasets and self-supervised models for sign language translation, outperforming all previous fully supervised approaches to this task. Our final model achieves highest BLEU-4 score, highest chrF score, and highest human evaluation score among all participants of the task. However, translation quality remains extremely low.

Acknowledgements

This work was supported in part by the TRI University 2.0 program. We thank Shester Gueuwou for helpful discussions about sign languages and translation.

References

Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. 2020. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? A new model and the Kinetics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. 2022. A simple multi-modality transfer learning baseline for sign language translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Konstantinos M. Dafnis, Evgenia Chroni, Carol Neidle, and Dimitri Metaxas. 2022. Bidirectional skeleton-based isolated sign recognition using graph convolution networks and transfer learning. In *13th International Conference on Language Resources and Evaluation (LREC)*.

Subhadeep Dey, Abhilash Pal, Cyrine Chaabani, and Oscar Koller. 2022. Clean text and full-body transformer: Microsoft's submission to the WMT22 shared task on sign language translation. *Proceedings of the Seventh Conference on Machine Translation (WMT)*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. 2020. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. 2023. Gloss-free end-to-end sign language translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Mathias Müller, Malihe Alikhani, Eleftherios Avramidis, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Sarah Ebling, Cristina España-bonet, Anne Goering, Roman Grundkiewicz, Mert Inan, Zifan Jiang, Oscar Koller, Amit Moryossef, Annette Rios, Dimitar Shterionov, Sandra Sidler-Miserez, Katja Tissi, and Davy Van Landuyt. 2023a. Findings of the second WMT shared task on sign language translation (WMT-SLT23). In *Proceedings of the Eighth Conference on Machine Translation (WMT)*.

Mathias Müller, Sarah Ebling, Eleftherios Avramidis, Alessia Battisti, Michèle Berger, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Cristina España-Bonet, Roman Grundkiewicz, et al. 2022. Findings of the first WMT shared task on sign language translation (WMT-SLT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*.

Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023b. Considerations for meaningful sign language machine translation based on glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Carol Neidle and Carey Ballard. 2022. Why alternative gloss labels will increase the value of the WLASL dataset. *ASL-LRP Project Report No. 21*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT)*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Marcelo Sandoval-Castañeda, Yanhong Li, Diane Brentari, Karen Livescu, and Gregory Shakhnarovich. 2023. Self-supervised video transformers for isolated sign language recognition. *arXiv preprint arXiv:2309.02450*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Bowen Shi, Diane Brentari, Greg Shakhnarovich, and Karen Livescu. 2022a. Open-domain sign language translation learned from online video. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Bowen Shi, Diane Brentari, Gregory Shakhnarovich, and Karen Livescu. 2022b. TTIC's WMT-SLT 22 sign language translation system. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022c. Learning audio-visual speech representation by masked multimodal cluster prediction. In *International Conference on Learning Representations (ICLR)*.

Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, and Donald Metzler. 2022. Scale efficiently: Insights from pretraining and finetuning transformers. In *International Conference on Learning Representations (ICLR)*.

David Uthus, Garrett Tanzer, and Manfred Georg. 2023. YouTube-ASL: A large-scale, open-domain american sign language-english parallel corpus. *arXiv preprint arXiv:2306.15162*.

Gül Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. 2021. Read and attend: Temporal localisation in sign language videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

R. Wang, D. Chen, Z. Wu, Y. Chen, X. Dai, M. Liu, Y. Jiang, L. Zhou, and L. Yuan. 2022. BEVT: BERT Pretraining of Video Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. 2021. Spatial-temporal multi-cue network for sign language recognition and translation. *IEEE Transactions on Multimedia*.

KnowComp Submission for WMT23 Sign Language Translation Task

Baixuan Xu[1], Haochen Shi[1], Tianshi Zheng[1], Qing Zong[2],
Weiqi Wang[1], Zhaowei Wang[1], Yangqiu Song[1]
[1]Department of Computer Science and Engineering, HKUST, Hong Kong SAR, China
[2]Harbin Institute of Technology (Shenzhen), Guangzhou, China
{bxuan, hshiah, tzhengad}@connect.ust.hk, zongqing0068@gmail.com
{wwangbw, zwanggy, yqsong}@cse.ust.hk

Abstract

Sign Language Translation (SLT) is a complex task that involves accurately interpreting sign language gestures and translating them into spoken or written language and vice versa. Its primary objective is to facilitate communication between individuals with hearing difficulties using deep learning systems. Existing approaches leverage gloss annotations of sign language gestures to assist the model in capturing the movement and differentiating various gestures. However, constructing a large-scale gloss-annotated dataset is expensive and impractical to cover multiple languages, and pre-trained generative models cannot be efficiently used due to the lack of textual source context in SLT. To address these challenges, we propose a gloss-free framework for the WMT23 SLT task. Our system primarily consists of a visual extractor for extracting video embeddings and a generator responsible for producing the translated text. We also employ an embedding alignment block that is trained to align the embedding space of the visual extractor with that of the generator. Despite undergoing extensive training and validation, our system consistently falls short of meeting the baseline performance. Further analysis shows that our model's poor projection rate prevents it from learning diverse visual embeddings. Our codes and model checkpoints are available at https://github.com/HKUST-KnowComp/SLT.

1 Introduction

Machine translation has significantly improved thanks to the development of pre-trained language models (Mohammadshahi et al., 2022; Huang et al., 2023). While translation within a single modality has been extensively studied, translation involving multiple modalities remains challenging and less explored (Lin et al., 2023). Sign Language Translation (SLT), which translates sign gestures into spoken language, remains an exceedingly complex task due to two fundamental challenges. Firstly, sign languages are visual-gestural languages that rely on manual signs, facial expressions, and body movements to convey information. This fundamental distinction sets them apart from written languages, which consist of word characters and symbols. Consequently, translation models must be able to accurately interpret visual signals and gestures and develop a deep understanding of the semantics involved in producing prompt translations. However, the multimodal nature of sign languages poses a significant challenge for models, requiring them to learn and generalize these complex interactions effectively. Moreover, sign languages are typically represented as exceedingly lengthy sequences of frames, surpassing the number of tokens in a standard sentence (Guo et al., 2018). This requires translation models to grasp the prolonged dependencies within the video to accurately capture the information conveyed through these visual signals.

To tackle these challenges, methods have been proposed that utilize pre-training a visual backbone based on gloss annotations (Camgöz et al., 2020). These approaches have demonstrated exceptional performance in various multimodal translation tasks. Nevertheless, the acquisition of extensive gloss annotations comes with significant cost and practical constraints, making it impractical to cover a wide range of multilingual translation directions (Müller et al., 2023).

In this paper, we propose a gloss-free framework for the SLT task. Our approach combines a pre-trained visual backbone model (Varol et al., 2021), which has been trained to recognize sign gestures, with a GPT2-based language model (Radford et al., 2019) to generate the translated sentence. To align the embedding space between both models, we utilize an embedding alignment block inspired by ClipCap (Mokady et al., 2021). The final translation is produced using converted visual embeddings and text embeddings (Section 3). Despite

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 351–358
December 6–7, 2023.

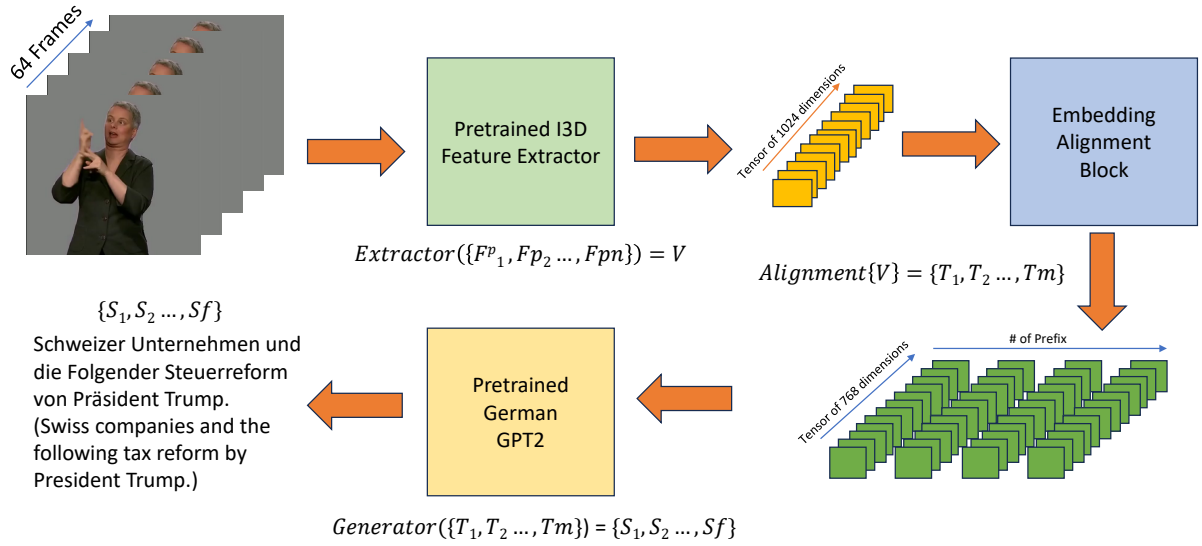


Figure 1: An overview of our framework. We first downsample video data and feed them into the visual feature extractor to obtain the visual embeddings. The embeddings are then passed into the alignment block to project them into embedding inputs of the `German-GPT2`. They are used as the prefix of the GPT2 model to generate the final translation results.

conducting extensive experiments with our system, we consistently achieved a BELU score of 0.1 and a chrF score of 7.6 on the testing set of the SRF dataset, which is below the baseline performance. Further analysis reveals that the embedding alignment block fails to differentiate between different embedding inputs from the visual encoder. As a result, our generation often produces repeated and nonsensical outputs. We will make all codes and results publicly available upon acceptance of this paper.

2 Preliminary

2.1 Task Definition

The objective of the Sign Language Translation (SLT) task (Fang et al., 2017; Kan et al., 2021) is to utilize the model's video understanding ability and language modeling ability to translate meaningful gesture sequence into spoken language (Varol et al., 2021; Hu et al., 2023). Formally, our objective is to learn a conditional probablity $P(S|F^r)$ of generating a natural spoken language, denoted as $\mathcal{S} = \{S_1, S_2..., S_m\}$ with m tokens given the raw sign language video $\mathcal{F}^r = \{F_1^r, F_2^r..., F_n^r\}$ with n frames.

To better elaborate our proposed model, we hereby set some notions for convenience. The aforementioned $\mathcal{S}$ and $\mathcal{F}^r$ refer to the translated spoken language and the sign language video before preprocessing. We use $\mathcal{F}^p = \{F_1^p, F_2^p..., F_n^p\}$ to denote the preprocessed video frames. In our proposed model, we endeavor to optimize the alignment block to yield better translation results while parameters in other modules are frozen for training efficiency.

Dataset	#raw data	#processed data
SRF	771	354901

Table 1: Statistics of the SRF dataset. # raw data refers to the number of videos, and # processed data is the amount of data after video slicing.

2.2 Dataset

We use the datasets provided by Müller et al. (2023) as our primary training and evaluation benchmarks. Our model is exclusively trained on the SRF dataset (Jiang et al., 2023b), while the SignSuisse dataset (Jiang et al., 2023a) is solely utilized for zero-shot evaluation purposes. Both datasets consist of sign language videos accompanied by their corresponding translation text in German. The statistical information for the SRF dataset can be

found in Table 1.

The SRF (Jiang et al., 2023b) dataset comprises videos from Standard German daily news (Tagesschau) and Swiss German weather forecast (Meteo) episodes broadcast. They are further interpreted into Swiss German Sign Language by hearing interpreters via Swiss National TV. In the SRF dataset released by Müller et al. (2022), there are a total of 354901 video slices covering episodes from 2014 to 2022.

The SignSuisse (Jiang et al., 2023a) dataset contains 18221 lexical items in Swiss German Sign Language, French Sign Language of Switzerland, and Italian Sign Language of Switzerland, represented as videos with corresponding spoken language translations.

The BSL-1k (Albanie et al., 2020) is a large-scale sign language recognition dataset constructed based on British Sign Language(BSL) signs. The authors leverage the observation that signers often mouth the word they are signing simultaneously, providing additional visual cues. They use visual keyword spotting to detect the mouthings and align them with the subtitles to determine whether and when a keyword of interest is uttered by a talking face using visual information. The dataset is then used to train a strong sign recognition model for co-articulated signs in BSL and serves as excellent pretraining for other sign languages and benchmarks. Thus, in our paper, it is reasonable for us to use a model pretrained on BSL-1k as our visual feature extractor and expect it to yield meaningful and informative video representations for the model to utilize.

3 Method

This section introduces our proposed framework, which is depicted in Figure 1. While previous systems (Dey et al., 2022; Shi et al., 2022; Tarres et al., 2022) primarily employ an encoder-decoder paradigm and train their models from scratch to address this task, we distinguish ourselves by being the first to utilize a pre-trained language model for this task, as these language models possess strong natural language understanding and generation ability (Wang et al., 2023c, 2022; Fang et al., 2021b,a, 2023; He et al., 2022; Bai et al., 2023a,b). Specifically, we leverage the pre-trained I3D model provided by Varol et al. (2021) as our visual extractor backbone and employ a German-GPT2 model (Schweter, 2020) as the generator's backbone.

3.1 Video Extractor

We use the Two-Stream Inflated 3D ConvNets (I3D; Carreira and Zisserman, 2017) that is pre-trained on the BSL-1k (Albanie et al., 2020) dataset as our visual extractor backbone. I3D was first proposed by Carreira and Zisserman (2017) aiming to mitigate the 2D convolution network failure to capture the temporal information behind the video data. To overcome this, I3D directly expands the original 2D convolution network, which yields significant success in 3-dimensional space by expanding extra dimension to the kernel and pooling layer. When the kernel and pooling layers are extended to 3D in I3D, these layers are initialized using the pre-trained weights from the corresponding 2D image classification networks. Overall, the I3D model offers a powerful framework for action recognition by leveraging the strengths of both image classification architectures and spatio-temporal feature extraction in videos. For the SLT task, we ask the model to transform a 64 frames ($\mathcal{F}_p$) video into a 1024-dimensional tensor ($\mathcal{V}$), denoted as:

$$\text{Extractor}(\{F_1^p, F_2^p ..., F_n^p\}) = \mathcal{V}$$

3.2 Embedding Alignment Block

Inspired by the success of ClipCap (Mokady et al., 2021), we then train an embedding alignment block to project the obtained visual embeddings $\mathcal{V}$ into textual embeddings T for further processing by German-GPT2. ClipCap was originally designed by Mokady et al. (2021) to tackle the task of image captioning (Ou et al., 2023). In the paper, the authors utilized the expressive power of an image feature extractor and a generative language model. By adding an alignment layer in between, the representation of the visual modality can be projected to the text modality for the language model to generate meaningful captions. The extraordinary ability shown by this innovative architecture makes it reasonable for us to adopt it in our framework. We implement the alignment block by stacking six transformer encoder layers together. Two fully connected neural networks are also placed before and after the alignment block to extend the visual embeddings into a sequential format and densify the aligned embeddings into prefix embeddings of German-GPT2, respectively. Formally, this process can be denoted as:

$$\text{Alignment}(\mathcal{V}) = \{T_1, T_2 ..., T_m\}$$

	BLEU			chrF			BLEURT		
Submission	all	SS	SRF	all	SS	SRF	all	SS	SRF
Baseline	0.09±0.03	0.15±0.06	0.10±0.05	12.4±0.4	12.2±0.5	12.5±0.5	0.072±0.003	0.083±0.005	0.060±0.005
KnowComp	0.07±0.05	0.06±0.02	0.11±0.09	7.6±0.3	8.2±0.4	7.2±0.4	0.083±0.005	0.084±0.007	0.081±0.007

Table 2: The experiment result of our proposed model comparing to the baseline released by the shared task organizer. Although our model was trained only on SRF, we still shown stronger performance on BLEURT than the baseline model in domain of SS and all. SS dataset is OOD and all is partially OOD for our model.

3.3 Text Generator

Finally, we leverage a pre-trained `German-GPT2` model as the text generator to generate the final translations by feeding the previously acquired textual prefix embeddings as the input. The German-GPT2 is trained on a large German corpus GC4 and can generate fluent german sentences. This step can be finally denoted as:

$$\text{Generator}(\{T_1, T_2..., T_m\} = \{S_1, S_2..., S_f\}$$

4 Experiments

4.1 Experiment Setup

We first describe our data preprocessing procedure and experiment settings.

4.1.1 Data Preprocessing

We first preprocess the raw data by dividing the video into smaller segments, or video slices, and match them with their corresponding ground truth German translations. To address a potential issue with the video extractor's encoding capacity, we adopt a downsampling strategy. Specifically, we select the first frame from every three frames in each video slice. Doing so reduces the number of frames and alleviates encoding challenges. Additionally, we encounter cases where certain video slices have fewer than 64 frames. To maintain consistency in video length, we append pure black frames to the end of these slices. To ensure compatibility with the video feature extractor's training environment, we resize each video frame to 224×224 dimension. This step guarantees that the model functions effectively within its designated parameters.

4.1.2 Experiment Setting

To enhance training efficiency, the parameters of the two backbone models are frozen, while the parameters of GPT2 are unfrozen after a certain iteration. This ensures that the randomly initialized transformer encoder does not compromise the language modeling ability of the GPT2 model. In our experiment, we set the batch size to 4, the learning rate to 5e-6, and changed the training parameters at iteration 66000. We employ an Adam (Kingma and Ba, 2015) as our optimizer and save the model checkpoint every 1000 iterations. The input and output lengths of GPT2 were fixed at 20, as we observed that most of the ground truth lengths were 20 or less, making this maximum length setting cover a significant portion of the training data. We set the number of heads in the multi-head attention to 8 and the prefix length for GPT2 to 4. Before feeding the embedding to the alignment block, the sequence length for translating the visual embedding was adjusted to 2×4, where 4 represents the GPT2 model's prefix number. Our model consists of 6 stacked encoder layers forming the alignment block. All experiments were conducted on NVIDIA GeForce GTX 1080 Ti with 11G memory.

4.2 Results

After extensive training and evaluation, our system achieves a BLEU (Papineni et al., 2002) score of 0.1 and a Chrf (Popovic, 2015) score of 7.6 in this shared task. These results are obtained from the official result submission platform. We present our experimental findings in comparison to the baseline model provided by the organizers, as shown in Table 2. Despite training our model solely on SRF, we outperform the baseline regarding the BLEURT (Sellam et al., 2020) score in SignSuisee and a combination of both datasets, which are considered out-of-domain evaluations for our model. However, it is important to note that our system falls significantly below the baselines and systems from other submissions. One potential explanation for this discrepancy could be that our system has not yet reached its optimal state, as the alignment block is trained from scratch, which could be quite challenging to converge. We conduct a fine-grained analysis in the following section to further investigate this hypothesis.

Original Subtitles	Generated Subtitles
Das Parlament muss nun auch die Städte ins Boot holen.<pad><pad><pad><pad><pad>	Der Schweiz. Sie werden in der Schweiz geboren. Deutschland.....
da fehlte oft das richtige Timing,<pad><pad><pad><pad><pad>	"Der Stadt Zürich. Zürich. Zürich. Zürich.. die Stadt Zürich. Zürich.. Zürich"
Am Samstagabend zunächst noch Föhn, dann wird es feuchter.<pad><pad><pad>	"Der Schweizer Regierungspräsidentin der Schweiz.. 20. 20. 20. 20. 20."
Danke, Andrea.<pad><pad><pad><pad>	"Der Film die Welt in den Abgrund. Rom. Deutschland....."
Es liegt an uns, Lösungen zu finden, um dieses Spiel zu gewinnen.<pad>	""Ich bin auch nicht, weil ich habe das nicht so viel. West.W.."

Table 3: Examples of our generated subtitles with their corresponding ground truth subtitles. We observe that 4 out of 5 of our generated sentences generate the same token for the first one and keep generating the same token at the end of its sentence. We try to analyze the reason for this in the following section.

4.3 Analysis

To analyze the reasons behind the failure of our system and its tendency to generate repetitive words in translations, we conduct a tSNE plot analysis of the visual embeddings before and after passing through the embedding alignment block. The results are presented in Figure 2. Upon examining the plot, we observe that the orange markers, representing the embeddings before alignment, were scattered, occupying a large area in the plot. In contrast, the blue crosses, corresponding to the embeddings after alignment, are densely concentrated in the middle of the plot. This stark contrast proves that the model loses its ability to differentiate between different visual features after projecting the embeddings from the I3D embedding space to the German-GPT2 embedding space. One potential explanation for this is that the embedding alignment block has not been effectively trained under the current training protocol. Further investigation is required to understand the underlying causes and devise appropriate solutions.

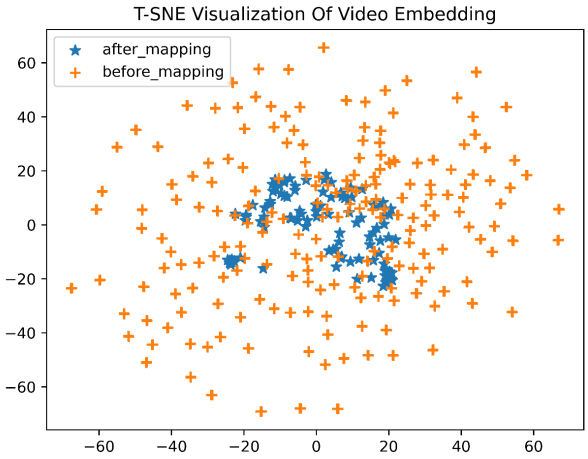


Figure 2: The tSNE comparison plot of the video embeddings before and after the embedding alignment block. We observe that the embeddings of different videos are dispersely distributed. However, they exhibit a denser distribution after alignment, which challenges generating coherent natural language descriptions.

4.4 Case Study

In Table 3, we present several instances of our generation using the data from the SRF dataset. The left column displays the ground truth sentences with a pad token appended at the end. In the right column, we showcase the generated sentences. Notably, 4 out of 5 of these sentences begin with "Der," and some consistently produce the same token, particularly in the final few positions. This further illustrates the subpar performance of our model. One possible explanation for this issue is the concentration of embeddings after the alignment block, which increases the likelihood of generating similar tokens. In the future, large-scale pertaining and appropriately leveraging large language models (OpenAI, 2023; Chan et al., 2023; Yu et al., 2023) and large multimodal foundation models (Zhu et al., 2023) may also be considered to improve the performance of this task further.

5 Conclusions

In conclusion, this paper presents the KnowComp system for the WMT23-SLT Sign Language Translation Shared Task. Our system utilizes two pretrained backbone models for visual feature extraction and translation text generation. However, this architecture fails, resulting in unsatisfactory perfor-

mance across all evaluation datasets. Our system's performance is significantly below the baseline's performance. We have identified a critical weakness in our model through further analysis, including embedding t-SNE plots and case studies. The embedding alignment block unexpectedly densifies all visual embeddings together, leading to the generator generating repeated tokens. To enhance our model's performance in future work, an appropriate data augmentation technique (Wang et al., 2023b,a; Gowda et al., 2022) can be implemented to help the alignment block distinguish different input features more efficiently. Also, future works can focus on whether further increasing the model capacity could help to mitigate the issue shown in the analysis section considering the advancing computation resources.

Acknowledgements

The authors would like to thank the committee of WMT2023, the organizers of the SLT task, and the anonymous reviewers. The authors of this paper were supported by the NSFC Fund (U20B2053) from the NSFC of China, the RIF (R6020-19 and R6021-20), and the GRF (16211520 and 16205322) from RGC of Hong Kong. We also thank the support from the UGC Research Matching Grants (RMGS20EG01-D, RMGS20CR11, RMGS20CR12, RMGS20EG19, RMGS20EG21, RMGS23CR05, RMGS23EG08).

References

Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. 2020. BSL-1K: scaling up co-articulated sign language recognition using mouthing cues. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI*, volume 12356 of *Lecture Notes in Computer Science*, pages 35–53. Springer.

Jiaxin Bai, Xin Liu, Weiqi Wang, Chen Luo, and Yangqiu Song. 2023a. Complex query answering on eventuality knowledge graph with implicit logical constraints. *CoRR*, abs/2305.19068.

Jiaxin Bai, Tianshi Zheng, and Yangqiu Song. 2023b. Sequential query encoding for complex query answering on knowledge graphs. *CoRR*, abs/2302.13114.

Necati Cihan Camgöz, Oscar Koller, Simon Hadfield, and Richard Bowden. 2020. Sign language transformers: Joint end-to-end sign language recognition and translation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10020–10030. Computer Vision Foundation / IEEE.

João Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? A new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4724–4733. IEEE Computer Society.

Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2023. Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations. *CoRR*, abs/2304.14827.

Subhadeep Dey, Abhilash Pal, Cyrine Chaabani, and Oscar Koller. 2022. Clean text and full-body transformer: Microsoft's submission to the WMT22 shared task on sign language translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 969–976, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Biyi Fang, Jillian Co, and Mi Zhang. 2017. Deepasl: Enabling ubiquitous and non-intrusive word and sentence-level sign language translation. In *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems, SenSys 2017, Delft, Netherlands, November 06-08, 2017*, pages 5:1–5:13. ACM.

Tianqing Fang, Quyet V. Do, Sehyun Choi, Weiqi Wang, and Yangqiu Song. 2023. CKBP v2: An expert-annotated evaluation set for commonsense knowledge base population. *CoRR*, abs/2304.10392.

Tianqing Fang, Weiqi Wang, Sehyun Choi, Shibo Hao, Hongming Zhang, Yangqiu Song, and Bin He. 2021a. Benchmarking commonsense knowledge base population with an effective evaluation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8949–8964. Association for Computational Linguistics.

Tianqing Fang, Hongming Zhang, Weiqi Wang, Yangqiu Song, and Bin He. 2021b. DISCOS: bridging the gap between discourse knowledge and commonsense knowledge. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 2648–2659. ACM / IW3C2.

Shreyank N. Gowda, Marcus Rohrbach, Frank Keller, and Laura Sevilla-Lara. 2022. Learn2augment: Learning to composite videos for data augmentation in action recognition. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXI*, volume 13691 of *Lecture Notes in Computer Science*, pages 242–259. Springer.

Dan Guo, Wengang Zhou, Houqiang Li, and Meng Wang. 2018. Hierarchical LSTM for sign language

translation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 6845–6852. AAAI Press.

Mutian He, Tianqing Fang, Weiqi Wang, and Yangqiu Song. 2022. Acquiring and modelling abstract commonsense knowledge via conceptualization. *CoRR*, abs/2206.01532.

Hezhen Hu, Weichao Zhao, Wengang Zhou, and Houqiang Li. 2023. Signbert+: Hand-model-aware self-supervised pre-training for sign language understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(9):11221–11239.

Kaiyu Huang, Peng Li, Jin Ma, Ting Yao, and Yang Liu. 2023. Knowledge transfer in incremental learning for multilingual neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15286–15304. Association for Computational Linguistics.

Zifan Jiang, Amit Moryossef, Mathias Müller, and Sarah Ebling. 2023a. Signsuisse dsgs/lsf/lis lexicon.

Zifan Jiang, Mathias Müller, Sarah Ebling, Amit Moryossef, and Robin Ribback. 2023b. Srf dsgs daily news broadcast: video and original subtitle data.

Jichao Kan, Kun Hu, Markus Hagenbuchner, Ah Chung Tsoi, Mohammed Bennamoun, and Zhiyong Wang. 2021. Sign language translation with hierarchical spatio-temporalgraph neural network. *CoRR*, abs/2111.07258.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kezhou Lin, Xiaohan Wang, Linchao Zhu, Ke Sun, Bang Zhang, and Yi Yang. 2023. Gloss-free end-to-end sign language translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 12904–12916. Association for Computational Linguistics.

Alireza Mohammadshahi, Vassilina Nikoulina, Alexandre Berard, Caroline Brun, James Henderson, and Laurent Besacier. 2022. Small-100: Introducing shallow multilingual machine translation model for low-resource languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 8348–8359. Association for Computational Linguistics.

Ron Mokady, Amir Hertz, and Amit H. Bermano. 2021. Clipcap: CLIP prefix for image captioning. *CoRR*, abs/2111.09734.

Mathias Müller, Malihe Alikhani, Eleftherios Avramidis, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Sarah Ebling, Cristina España-bonet, Anne Goering, Roman Grundkiewicz, Mert Inan, Zifan Jiang, Oscar Koller, Amit Moryossef, Annette Rios, Dimitar Shterionov, Sandra Sidler-miserez, and Katja Tissi. 2023. Findings of the second WMT shared task on sign language translation (WMT-SLT23). In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Mathias Müller, Sarah Ebling, Eleftherios Avramidis, Alessia Battisti, Michèle Berger, Richard Bowden, Annelies Braffort, Necati Cihan Camgöz, Cristina España-Bonet, Roman Grundkiewicz, Zifan Jiang, Oscar Koller, Amit Moryossef, Regula Perrollaz, Sabine Reinhard, Annette Rios, Dimitar Shterionov, Sandra Sidler-Miserez, and Katja Tissi. 2022. Findings of the first WMT shared task on sign language translation (WMT-SLT22). In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 744–772. Association for Computational Linguistics.

Mathias Müller, Zifan Jiang, Amit Moryossef, Annette Rios, and Sarah Ebling. 2023. Considerations for meaningful sign language machine translation based on glosses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 682–693. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Jiefu Ou, Benno Krojer, and Daniel Fried. 2023. Pragmatic inference with a CLIP listener for contrastive captioning. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1904–1917. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Maja Popovic. 2015. chrf: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation, WMT@EMNLP 2015, 17-18 September 2015, Lisbon, Portugal*, pages 392–395. The Association for Computer Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Stefan Schweter. 2020. German gpt-2 model.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7881–7892. Association for Computational Linguistics.

Bowen Shi, Diane Brentari, Gregory Shakhnarovich, and Karen Livescu. 2022. TTIC's WMT-SLT 22 sign language translation system. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 989–993, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Laia Tarres, Gerard I. Gállego, Xavier Giro-i nieto, and Jordi Torres. 2022. Tackling low-resourced sign language translation: UPC at WMT-SLT 22. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 994–1000, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Gül Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. 2021. Read and attend: Temporal localisation in sign language videos. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 16857–16866. Computer Vision Foundation / IEEE.

Weiqi Wang, Tianqing Fang, Wenxuan Ding, Baixuan Xu, Xin Liu, Yangqiu Song, and Antoine Bosselut. 2023a. CAR: conceptualization-augmented reasoner for zero-shot commonsense question answering. *CoRR*, abs/2305.14869.

Weiqi Wang, Tianqing Fang, Baixuan Xu, Chun Yi Louis Bo, Yangqiu Song, and Lei Chen. 2023b. CAT: A contextualized conceptualization and instantiation framework for commonsense reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13111–13140. Association for Computational Linguistics.

Zhaowei Wang, Quyet V. Do, Hongming Zhang, Jiayao Zhang, Weiqi Wang, Tianqing Fang, Yangqiu Song, Ginny Y. Wong, and Simon See. 2023c. COLA: contextualized commonsense causal reasoning from the causal inference perspective. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5253–5271. Association for Computational Linguistics.

Zhaowei Wang, Hongming Zhang, Tianqing Fang, Yangqiu Song, Ginny Y. Wong, and Simon See. 2022. Subeventwriter: Iterative sub-event sequence generation with coherence controller. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1590–1604. Association for Computational Linguistics.

Changlong Yu, Weiqi Wang, Xin Liu, Jiaxin Bai, Yangqiu Song, Zheng Li, Yifan Gao, Tianyu Cao, and Bing Yin. 2023. Folkscope: Intention knowledge graph construction for e-commerce commonsense discovery. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1173–1191. Association for Computational Linguistics.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592.

A Fast Method to Filter Noisy Parallel Data WMT2023 Shared Task on Parallel Data Curation

Minh-Cong Nguyen-Hoang[1] **Vinh Nguyen Van**[2] **Le-Minh Nguyen**[1]

[1]*Japan Advanced Institute of Science and Technology, JAIST*
[2]*University of Engineering and Technology, VNU, Hanoi, Vietnam*
{congnhm, nguyenml}@jaist.ac.jp
vinhnv@vnu.edu.vn

Abstract

The effectiveness of a machine translation (MT) system is intricately linked to the quality of its training dataset. In an era where websites offer an extensive repository of translations such as movie subtitles, stories, and TED Talks, the fundamental challenge resides in pinpointing the sentence pairs or documents that represent accurate translations of each other. This paper presents the results of our submission to the shared task WMT2023 (Sloto et al., 2023), which aimed to evaluate parallel data curation methods for improving the MT system. The task involved alignment and filtering data to create high-quality parallel corpora for training and evaluating the MT models. Our approach leveraged a combination of dictionary and rule-based methods to ensure data quality and consistency. We achieved an improvement with the highest 1.6 BLEU score compared to the baseline system. Significantly, our approach showed consistent improvements across all test sets, suggesting its efficiency.

1 Introduction

Neural Machine Translation (NMT) has revolutionized the field of machine translation by utilizing deep learning algorithms to learn from large amounts of data and generate high-accurate translations (Sennrich et al., 2016; Vaswani et al., 2017). However, the success of NMT models heavily depends on the quantity and quality of data used for training. On low-resource language pairs, the NMT architectures perform poorly (Koehn and Knowles, 2017; Khayrallah and Koehn, 2018) and are more sensitive to noisy data than statistical machine translation (SMT) methods (Belinkov and Bisk, 2017; Koehn et al., 2018). Therefore, access to vast cleaned corpus can significantly improve the performance of NMT models, allowing them to learn and produce more accurate translations (Bojar et al., 2017).

Fortunately, very large text sources offer a massive collection of data for various types of content, including movie subtitles, stories, and TED Talks. These resources have not been fully exploited for NMT training due to the lack of alignment between the source and target languages. Furthermore, the parallel data which movie subtitles also could be noisy with poor accuracy (Khayrallah and Koehn, 2018). To address this challenge, WMT2023 introduced a shared task on Parallel Data Curation for the Estonian-Lithuanian (et-lt) language pair, focusing on finding the best possible training data set within the web-crawled data to train a downstream NMT model (Sloto et al., 2023).

Among the popular solutions, Thompson and Koehn (2019) introduced using Vecalign to embed sentences and compute the cosine similarity of sentence pairs. Following this method, the shared task provides participants with extensive cosine similarity files and LASER embeddings generated by the LASER model (Heffernan et al., 2022). Participants are tasked with identifying the most optimal parallel data to train the MT models. Although this approach performs efficiently in many cases, Zhou et al. (2022) has shown that the cosine similarity has several limitations. Because the sentence representation in vector space could be impacted by word frequency. To tackle this problem, we build a pipeline to improve the quality of the parallel corpus. Our contributions focus on:

- using the phrased base dictionary to distill the high-quality sentences.
- making the pipeline to re-ranking the top-K cosine similarity.
- analyzing the influence of cosine similarity thresholds on corpus size and MT Models.

The related work is presented in section 2. The detail of our method is described in section 3, experi-

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 359–365
December 6–7, 2023.

ments and results are shown in section 4. Finally, the analysis is presented in section 4.5.

2 Related work

The WMT2023 shared task builds upon previous shared tasks focused on document alignment (WMT 16) and sentence filtering (WMT 18, 19, 20) (Buck and Koehn, 2016; Koehn et al., 2018, 2019, 2020). Previously, several researchers proposed a method to align documents, such as Gomes and Pereira Lopes (2016) used the phrase table to align in the search space and then fill in and refine alignments. Moreover, Thompson and Koehn (2019) employed the Vecalign to gain the sentence embeddings. Nevertheless, Sentence alignments based on cosine similarity have some limitations because the cosine scores could be dense in the range of 0.5 to 1 (Zhou et al., 2022). And with the same query sentence, the higher score could not determine the quality of the parallel sentence.

In addition, for the filtering shared task, participants applied filtering rules to eliminate noisy data, including removing too long/short sentences, using language identification for source and target (Kejriwal and Koehn, 2020) or fine-tuning pre-trained models such as BERT, XLM to re-score sentence pairs (Yang et al., 2019; Bernier-Colborne and Lo, 2019; Açarçiçek et al., 2020). Besides, Xu and Koehn (2017) created artificially noisy data by generating inadequate and nonfluent translations. They used this noisy data to train a classifier to distinguish between high-quality and low-quality sentence pairs within a corpus containing noise.

We found the related ideas from (Lu et al., 2020; Xu et al., 2020). Both of these approaches only focus on the alignment rules and adopt the other pre-trained models. Junczys-Dowmunt (2018) trained an NMT model to filter data and became standard for the high-resource case. Nevertheless, when training an original NMT model with low or noisy resources, the NMT model could face certain limitations. In our work, we utilize the phrase table to compute *edit distance* and extract the superior sentences. Furthermore, we introduce a pipeline to re-rank sentences based on their top-K cosine similarity scores and extract the best compact corpus for training purposes. The detail of our method is presented in section 3.

3 Methodology

3.1 LASER Similarity Scores

The LASER2 similarity scores are produced for the WMT23 shared task. These files are an intermediate output from our baseline submission. The laser embeddings applied L2 normalization and added them to a flat inner product index, such that the resulting scores are equivalent to cosine similarity. And query each index with all L2 normalized embeddings in the target sentences and store the top-8 results (locally, per chunk). Finally, the data is aggregated and meticulously sorted across unique IDs.

3.2 Building Dictionary

Our proposed method incorporates several innovative techniques to enhance the accuracy and efficacy of the filtering process. Initially, we train a phrased table based on MGiza++[1] (Gao and Vogel, 2008), a widely utilized algorithm for learning phrase tables from parallel corpora. Given a source string $X^I = \{x_1...x_i...x_I\}$ and a target string $Y^J = \{y_1...y_j...y_J\}$. In the context of statistical alignment, the probability of a source sentence given a target sentence is formulated as follows:

$$P(X^I|Y^J) = \sum_{i=1}^{J} P_\theta(X_i^I, a_i^J|Y_i^J), \qquad (1)$$

Where a_i^J represents the alignment of the sentence pair. The parameters θ can be estimated using maximum likelihood estimation (MLE) on a training corpus to represent the statistical probability with the best alignment of the sentence pair:

$$a_1^J = \arg\max_{a_1^J} p_\theta(x_i^I, a_i^J|y_i^J), \qquad (2)$$

These steps enable us to establish connections between words in the source and target languages. After that, we extract the dictionary from the phrase base table. This stage helps to remove unnecessary or redundant words and sentences, streamlining the dictionary and improving its quality.

3.3 Edit Distance

In this work, we utilized the dictionary to translate source sentences to target sentences called candidate strings. To identify sentences where the source and target are similar, we compute the edit

[1] https://github.com/moses-smt/mgiza

distance between a pair of candidate and reference sentences.

$$score = \sigma(C_i^I, Y_i^J), \quad (3)$$

Here, C_i^I is the candidate sentence and Y_i^J indicates the target sentence. The σ function employs the Damerau-Levenshtein distance (Miller et al., 2010). We set each insertion, deletion, and substitution as one step, but the transposition (swapping) of two words is computed as $\frac{1}{2}$ step. We opted for option $\frac{1}{2}$ in the swapping step due to the limitation of using a dictionary to translate strings, which neglects word positions. Finally, we choose the sentences that have scores greater than or equal to $\frac{N}{2}$, with N being the max length of the candidate sentence and target sentence.

3.4 Accumulative Filtering

Because the word frequency impacts the cosine score, we produce a filtering pipeline to improve the corpus quality and apply it to the larger corpus. The implementation method is described in Algorithm 1.

Algorithm 1: Accumulative Filtering

Data: The raw parallel corpora: $(X^I, Y^J) \in D$, dictionary, and threshold_values in range {0.7-0.9}
Result: The cleaned data: $(X_f, Y_f) \in D$
`/* Initialize the NMT model` θ `*/`
$t \leftarrow 0.9$;
$X, Y \leftarrow filter(D, t)$;
`/* Filter data via top1 cosine score with the threshold of t */`
$C' \leftarrow translate(X, dictionary)$;
`/* Translate source using dictionary */`
$X_f, Y_f \leftarrow select(C', Y)$;
`/* Select sentences based on edit-distance score */`
$\theta \leftarrow train(X_f, Y_f)$;
`/* Loop t with the step as 0,5 */`
for $t \in threshold\_values$ **do**
  $X, Y \leftarrow filter\_topK(D, t, 8)$;
  $X' \leftarrow translate(X, \theta)$;
  $X_f, Y_f \leftarrow select(C', Y)$;
  $\theta \leftarrow train(X_f, Y_f)$;
end

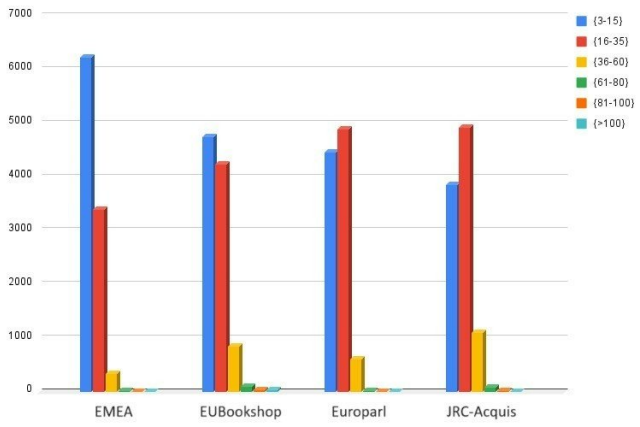


Figure 1: The statistics of the sentence length of each test set are used to evaluate the cleaned corpus. We separate the length of sentences into five levels, with 3-15 as the total sentences exhibit lengths that fall within the 3 to 5 range. The 16-35, 36-60, 61-80, 81-100, and >100 are the 16 to 35, 36 to 60, 61 to 80, 81 to 100, and greater 100 correspondingly.

4 Experiments

In this section, we describe the experimental setup for our system, including the data, training tools, and baseline system.

4.1 Data

In this shared task, the corpus was gathered from a recent snapshot of CommonCrawl[2].

Training Data

From the crawled data, the data is smoothed to some steps such as extracting plain texts from HTML documents, using the identifier language to hold the Estonian and Lithuanian documents, and removing the unsafe and offensive content. Besides, each sentence is assigned a distinct, randomly generated unique ID. These identifiers are uniform within their language datasets but diverge between two languages. This allows quick access and operation with data. Table 2 depicts the total number of collected data.

Testing Data

To assess the quality of the cleaned corpus, we train NMT models and evaluate them in four test sets, including EMEA, EU-Bookshop, Europarl, and JRC-Acquis. The statistics of the sentence length of each test set are exhibited in Figure 1, with the

[2]https://commoncrawl.org/blog/jan-feb-2023-crawl-archive-now-available

	EMEA BLEU	EU-Bookshop BLEU	Europarl BLEU	JRC-Acquis BLEU	EMEA chrF	EU-Bookshop chrF	Europarl chrF	JRC-Acquis chrF
LASER (Baseline)	18.3	19.1	18.1	24.3	49.7	52.3	51.8	55.2
Dictionary +Edit-Dist	18.3	19.1	18.5	24.3	49.7	52.3	51.8	54.9
Accummulative Filtering:								
Threshold-0.9	18.1	20.0	18.3	25.1	49.6	52.2	51.9	55.1
Threshold-0.85	18.5	**20.3**	19.1	25.4	49.7	52.7	**52.2**	55.2
Threshold-0.8*	**19.2**	20.1	**19.2**	25.7	**49.9**	**52.8**	52.1	55.4
Threshold-0.75	19.0	20.2	18.9	**25.9**	49.8	52.7	52.0	**55.6**

Table 1: The evaluation of BLEU scores and chrF scores for the filtering and alignment corpus.

No.	Estonian	Lithuanian
Num of Sents	53,279,844	63,556,320

Table 2: The statistics of sentences are available in the corpus for the Estonian-Lithuanian.

total number of sentences in each test set being 10,000.

4.2 Training Tools

We utilize the training scripts[3] provided by organizers to run the evaluation for the Shared Task. To observe the effect of filtered datasets, we use the same hyper-parameters for the whole experiment to compare results equally. In more detail, the Transformer architecture (Vaswani et al., 2017) is used in the training tool with the default 8 heads, 6 layers, and the model size is 512. Besides, the training pipeline employs the subword segmentation tool provided by (Sennrich et al., 2016) for tokenization. We use the sacreBLEU (Post, 2018) and ChrF++ (Popović, 2015) score to evaluate whole experiments.

4.3 Baseline

Following scripts provided by organizers, we present briefly how to create a simple baseline. Firstly, we collect the whole provided cosine similarity files. Secondly, we extract sentence alignments with the threshold of 0.9 and only select the top highest similarity scores. And finally, we run the end-to-end evaluation to produce BLEU scores from the extracted data.

[3]https://github.com/awslabs/sockeye/tree/wmt23_data_task

4.4 Our system

In the first place, we obtain the cosine files that are computed from Laser embeddings. From these files, We extract the sentence pairs by considering the highest cosine similarity score, specifically the top-1 score, and we set a threshold of 0.9. In the following phase, we remove longer sentences having 200 tokens and more and utilize the dictionary to perform word-by-word translation of these source sentences into the target language. After that, we compute the edit distance and eliminate poor-quality sentences. Finally, we employ the cumulative filtering algorithm discussed in section 3.4 to acquire the expanded corpus, opting for thresholds of 0.9, 0.85, and 0.8, respectively.

4.5 Results

In this section, we present the results obtained through a comparative analysis of different methods within the context of our works. Table 1 illustrates the results attained in the development system while preparing the submission. The system responsible for generating the scores for our final submission is shown in underline. We consider the reported results as the LASER baseline, and the outperforming results are indicated in bold.

Our investigation reveals that the LASER baseline provides a starting point for evaluation and moderate levels of performance across a range of metrics. However, the Accumulative Filtering approach, particularly when applying lower threshold values (0.85, 0.8, and 0.75), showcases significant improvements in various metrics. Notably, the choice of threshold within the Accumulative Filtering method influences performance, with lower thresholds yielding higher results. These findings

	top1_0.95	top1_0.9	top1_0.85	top1_0.8	top1_0.75	top1_0.7
Corpus size	173,239	1,230,810	4,194,132	12,918,719	27,811,424	32,568,712
BLEU	16.8	23.9	25.1	25.4	25.0	24.4

Table 3: The influence of sentences on the corpus size and BLEU score. The evaluation of the BLEU score is conducted specifically on the JRC-Acquis test set. Sentence selection is based solely on the cosine score threshold, with additional criteria involving the removal of excessively short or long sentences.

underscore the importance of threshold selection and methodological considerations in achieving optimal outcomes. Further analysis and task-specific considerations are required to determine the most suitable approach for our specific research objectives. We analyze the impact of the cosine similarity score thresholds on the corpus size and quality of NMT models. The details are described in section 4.5.

5 Analysis

In this section, we delve deeply into our approaches and the scale of our data corpora. Firstly, we conduct some experiments to find the best threshold when selecting the top-K highest cosine similarity score. For every source sentence, our approach involves selecting a single target sentence from a set of eight candidates based on the highest cosine similarity score provided. Table 3 illustrates the impact of different sentence selection criteria, denoted by the cosine similarity thresholds (top1_0.95, top1_0.9, top1_0.85, top1_0.8, top1_0.75, top1_0.7), on both the corpus size and BLEU score. The corpus size varies significantly depending on the threshold, ranging from 173,239 sentences to 32,568,712 sentences. Simultaneously, the BLEU score, evaluated on the JRC-Acquis test set, fluctuates, with the highest score of 25.4 achieved at the top1_0.8 threshold. These findings underscore the delicate balance between corpus size and translation quality, highlighting the importance of threshold selection in the context of machine translation evaluation.

Secondly, we conduct a statistical analysis to determine the number of sentence pairs that achieve the highest cosine score but are not considered parallel sentences. Table 4 shows the statistics for the number of sentences that do not have the highest cosine similarity score but are regarded as parallel sentences. The table indicates a total of 5,981,148 sentences in cleaned data and 353,642 sentences are considered re-ranking parallel sentences.

No.	Cleaned Data	Re-ranking
Num of Sents	5,981,148	353,642

Table 4: Number of sentences that are not in top-1 cosine similarity score, but are considered parallel sentences.

6 Conclusion

In conclusion, our study has provided valuable insights into the performance of different methods employed in our research on WMT2023 parallel data curation shared tasks. Our findings reveal that while the LASER baseline and using the dictionary method exhibited moderate and consistent performance across several metrics, the accumulative filtering approach, particularly when adopting lower threshold values (0.85, 0.8, and 0.75), demonstrated notable improvements in various aspects. Notably, the selection of the threshold played a pivotal role in influencing performance outcomes. Furthermore, our analysis also encompassed the identification of sentence pairs that exhibit parallel characteristics, even if they may not always possess the highest cosine similarity scores. In the future, further investigation and task-specific considerations will be essential in finding the smallest possible set of training data and achieving the highest result.

References

Haluk Açarçiçek, Talha Çolakoğlu, Pınar Ece Aktan Hatipoğlu, Chong Hsuan Huang, and Wei Peng. 2020. Filtering noisy parallel corpus using transformers with proxy task learning. In *Proceedings of the Fifth Conference on Machine Translation*, pages 940–946, Online. Association for Computational Linguistics.

Yonatan Belinkov and Yonatan Bisk. 2017. Synthetic and natural noise both break neural machine translation.

Gabriel Bernier-Colborne and Chi-kiu Lo. 2019. NRC parallel corpus filtering system for WMT 2019. In *Proceedings of the Fourth Conference on Machine*

Translation (Volume 3: Shared Task Papers, Day 2), pages 252–260, Florence, Italy. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Christian Buck and Philipp Koehn. 2016. Findings of the WMT 2016 bilingual document alignment shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 554–563, Berlin, Germany. Association for Computational Linguistics.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio. Association for Computational Linguistics.

Luís Gomes and Gabriel Pereira Lopes. 2016. First steps towards coverage-based document alignment. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 697–702, Berlin, Germany. Association for Computational Linguistics.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages.

Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

Ankur Kejriwal and Philipp Koehn. 2020. An exploratory approach to the parallel corpus filtering shared task WMT20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 959–965, Online. Association for Computational Linguistics.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742, Online. Association for Computational Linguistics.

Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino. 2019. Findings of the WMT 2019 shared task on parallel corpus filtering for low-resource conditions. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 54–72, Florence, Italy. Association for Computational Linguistics.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the WMT 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 726–739, Belgium, Brussels. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Jun Lu, Xin Ge, Yangbin Shi, and Yuqi Zhang. 2020. Alibaba submission to the WMT20 parallel corpus filtering task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 979–984, Online. Association for Computational Linguistics.

F.P. Miller, A.F. Vandome, and J. McBrewster. 2010. *Damerau-Levenshtein Distance*. Alphascript Publishing.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting bleu scores.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units.

Steve Sloto, Brian Thompson, Huda Khayrallah, Tobias Domhan, Thamme Gowda, and Philipp Koehn. 2023. Findings of the WMT 2023 shared task on parallel data curation. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore. Association for Computational Linguistics.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Hainan Xu and Philipp Koehn. 2017. Zipporah: a fast and scalable data cleaning system for noisy web-crawled parallel corpora. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2945–2950, Copenhagen, Denmark. Association for Computational Linguistics.

Runxin Xu, Zhuo Zhi, Jun Cao, Mingxuan Wang, and Lei Li. 2020. Volctrans parallel corpus filtering system for WMT 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 985–990, Online. Association for Computational Linguistics.

Yinfei Yang, Gustavo Hernandez Abrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax.

Kaitlyn Zhou, Kawin Ethayarajh, Dallas Card, and Dan Jurafsky. 2022. Problems with cosine as a measure of embedding similarity for high frequency words. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 401–423, Dublin, Ireland. Association for Computational Linguistics.

A Sentence Alignment Approach to Document Alignment and Multi-faceted Filtering for Curating Parallel Sentence Pairs from Web-crawled Data

Steinþór Steingrímsson
The Árni Magnússon Institute for Icelandic Studies
Reykjavik, Iceland
steinthor.steingrimsson@arnastofnun.is

Abstract

This paper describes the AST submission to the WMT23 Shared Task on Parallel Data Curation. We experiment with two approaches for curating data from the provided web-scraped texts. We use sentence alignment to identify document alignments in the data and extract parallel sentence pairs from the aligned documents. All other sentences, not aligned in that step, are paired based on cosine similarity before we apply various different filters. For filtering, we use language detection, fluency classification, word alignments, cosine distance as calculated by multilingual sentence embedding models, and Bicleaner AI. Our best model outperforms the baseline by 1.9 BLEU points on average over the four provided evaluation sets.

1 Introduction

The aim of the Shared Task on Parallel Data Curation at the Eighth Conference on Machine Translation (WMT23) is to evaluate parallel data curation methods (Sloto et al., 2023). The goal is to find the best machine translation (MT) training data within a provided pile of web-crawled data.

The language pair chosen for the task is Estonian-Lithuanian. The provided data is extracted from a single snapshot of CommonCrawl,[1] which according to the task organizers should contain enough training data to train a reasonable Estonian → Lithuanian MT model, even with limited compute. As well as providing the data, the organizers release pre-computed intermediate steps from a baseline, so participants can choose whether to focus on one or more aspects of the task. We describe the provided data and the baseline in Section 3.

In our submission we experiment on two aspects of parallel data curation. Initially we try to to identify parallel documents in the two languages. We then align sentences in the documents using our own sentence alignment tool, SentAlign[2] (Steingrímsson, 2023; Steingrímsson et al., 2023b), and train an MT system on the resulting sentence pairs. SentAlign is a sentence aligner that uses LaBSE (Feng et al., 2022) to score all possible alignment combinations for a document pair, selects the highest scoring one, but then re-evaluates the results by looking at each individual alignment and their closest neighbours to see if localized scores can be raised. This is to counteract an effect of dynamic programming with cosine similarity, which often favours many-to-many alignments over 1-to-1 alignments (see e.g. Thompson and Koehn (2019). Steingrímsson et al. (2023b) show that this approach outperforms other aligners on two evaluation sets, as well as on a downstream task. The other aligners include aligners such as the length based Gale-Church (Gale and Church, 1991), MT-based Bleualign (Sennrich and Volk, 2010) and Vecalign (Thompson and Koehn, 2019) which is the most similar to SimAlign, using LASER embeddings (Artetxe and Schwenk, 2019b) to calculate cosine similarity of alignment candidates, and a recursive approximation to reduce the search space, as opposed to evaluating all possibilities as SentAlign does. We describe our approach to document alignment in Section 4.1. Subsequently, we try to identify parallel sentence pairs in all the other provided data and run a number of different filters to remove sentence pair candidates that we deem likely to be detrimental or useless for MT training. Our filtering approaches are described in Section 4.2

2 Related Work

Khayrallah and Koehn (2018) show that incorrect translations, untranslated target text, misalignments, and other noisy segments in a parallel corpus have a detrimental effect on the output quality of neural machine translation (NMT) systems trained on that corpus, as measured by using BLEU (Papineni et al., 2002). They specify five general

[1] https://commoncrawl.org/
[2] https://github.com/steinst/SentAlign

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 366–374
December 6–7, 2023.

classes of noise commonly found in a German-English version of the ParaCrawl corpus: misaligned sentences, disfluent text, wrong language, short segments, and untranslated sentences. They find this distinction to be useful to give a general idea of which types of errors seem to have the least impact on MT systems (short segments, untranslated source sentences and wrong source language) and which have the most dramatic effect (untranslated target sentence). Misalignments, misordered words, and wrong language, in source or target texts, are also shown to be harmful, but not as harmful.

The Conference on Machine Translation, WMT, hosted three annual shared tasks on parallel corpus filtering (Koehn et al., 2018, 2019, 2020), focusing on filtering noisy web-crawled corpora. Submitted systems include the ones by Chaudhary et al. (2019) and Artetxe and Schwenk (2019a), who introduce approaches based on cross-lingual sentence embeddings trained from parallel sentences. Both papers use cosine similarity and consider the margin between a given sentence pair and its closest candidates to normalize the similarity scores.

Bicleaner (Sánchez-Cartagena et al., 2018; Ramírez-Sánchez et al., 2020) uses a set of handcrafted rules to detect flawed sentences and then proceeds to use a random forest classifier based on lexical translations and several shallow features such as respective length, matching numbers and punctuation. It also scores sentences based on fluency using 5-gram language models. The tool ranked highly on the first two parallel corpus filtering tasks at WMT. Bicleaner AI (Zaragoza-Bernabeu et al., 2022) is a fork of Bicleaner using a neural classifier. It has been shown to give significant improvements in translation quality as measured by BLEU when used for filtering training data for multiple language pairs, as compared to the previous version of the tool.

In contrast to tools that apply a single method for parallel corpus filtering, Aulamo et al. (2020) implement multiple different filters in the OpusFilter toolbox. These include heuristic based filters, language identification, character-based language models and word alignment tools. The toolbox can furthermore be extended with custom filters.

Herold et al. (2022) revisit the noise classes specified by Khayrallah and Koehn (2018) to investigate how accurately two strong filtering approaches, cross entropy (Rossenbach et al., 2018) and LASER (Artetxe and Schwenk, 2019b) can filter out different noise classes. They find that for a common language pair, German→English, most types of noise can be detected with over 90% accuracy, although misalignments and poor synthetic translation can only be detected with an accuracy of less than 70%. For a less common language pair, Khmer–English, the performance of the filtering system degraded significantly and the accuracy of identifying noise was lowered by 8–19%, depending on the type of noise, indicating that results can vary dramatically depending on the languages.

3 Data and Baseline

The provided data is retrieved from the 2023-06 snapshot of Common Crawl. The organizers have extracted plain text from HTML documents and used the Fasttext (Joulin et al., 2017) language identification model to remove documents not classified as Estonian or Lithuanian by the model, based on the first 2,000 characters of the document. Unsafe and offensive content has been removed. Documents from host names in the following lists in the blocklist project[3] where removed: *abuse, basic, crypto, drugs, fraud, gambling, malware, phishing, piracy, porn, ransomware, redirect, scam, torrent*. Documents were split into paragraphs based on line breaks, and then into sentences using Mediacloud Sentence Splitter.[4] Each sentence was assigned a unique sentence id and classified using the Fasttext language identification model. The data is provided in TSV format.

The task organizers provide LASER2 sentence embeddings (Heffernan et al., 2022) for all sentences in the correct language, as classified by the Fasttext model. They index the embeddings and query each index to retrieve the top-8 results for each sentence, based on cosine similarity. We use these results as a starting point for our filtering approaches, as described in Section 4.2. The baseline simply takes top-1, i.e. the highest scoring sentence pair for each sentence, provided the score exceeds a threshold of 0.9. This results in a set of 2,654,090 parallel pairs for training a baseline model.

A script for training the baseline model using Sockeye (Hieber et al., 2022) was provided. We use the script and Sockeye for training all models on a single Nvidia GeForce RTX 3090 GPU card.

[3] https://github.com/blocklistproject/Lists

[4] https://github.com/mediacloud/sentence-splitter

4 System Architecture

In this section we describe our approaches to the parallel data curation problem. First, we try to identify parallel documents in the two languages and align them on the sentence level. Second, we use the provided sentence pair candidates, eight for each sentence in each language, and filter using a number of different approaches to remove possibly detrimental pairs from our training set. The sentence pairs from the aligned documents and the filtered sentence pairs are combined to compile our final dataset.

4.1 Document identification by Sentence Alignment

Bilingual document alignment is a matching task that considers documents in two languages and estimates the likelihood of the documents being a translation of each other. In the Bilingual Document Alignment Shared Task at WMT 2016 (Buck and Koehn, 2016), the submitted systems used a variety of approaches. Some of these include Gomes and Pereira Lopes (2016), who used a phrase table from a phrase-based statistical machine translation (SMT) system to compute coverage scores. Dara and Lin (2016) use MT to find corresponding documents based on n-gram matches, assisted by document length ratio, and Mahata et al. (2016) use text matching based on sentence alignment and word dictionaries. Thompson and Koehn (2020) present a document alignment method that uses information on sentence order both when generating candidates and when re-scoring the candidates. For re-scoring the candidate pairs they perform sentence alignment and score the alignment based on semantic similarity of the resulting sentence pairs.

In this paper, we use sentence alignment and average cosine distance as measured by LaBSE (Feng et al., 2022) to determine whether documents can be aligned. The provided dataset contains sentences scraped from the web, information on the web domain and an order of sentences within documents on the websites. We recreate documents, most likely to have a corresponding translation in the other language, using this information. In order to reduce the need for compute we only consider texts from the same domain to be possible candidates for document alignment.

Our approach is the following:

1. We start by collecting a list of all web domains common to both languages.
2. From these domains, we recreate all documents that contain more than five sentences. The recreated documents have one sentence in each line.
3. Using SentAlign, for each domain we align the recreated documents in Estonian to all the recreated documents in Lithuanian, and vice versa. SentAlign outputs all aligned sentence pairs, as well as the LaBSE score for the pair.
4. If more than half of the sentences in either language does not get an alignment, the document pair is discarded.
5. If the average LaBSE score for all sentence alignments for a given document pair is below a threshold of 0.6, the document pair is discarded.
6. We calculate an alignment ratio using Equation 1:

$$\frac{1}{2}\left(\frac{et_{aligned}}{et_{total}} + \frac{lt_{aligned}}{lt_{total}}\right) \quad (1)$$

Where $et_{aligned}$ is the number of Estonian sentences that obtain an alignment to a Lithuanian sentence, and et_{total} is the total number of Estonian sentences in the document. $lt_{aligned}$ and lt_{total} are the corresponding numbers for Lithuanian.

From a pool of documents for each web domain, a greedy algorithm selects the document pair with the highest alignment ratio, and if multiple pairs have the highest ratio, one of those with the highest average LaBSE score. The selected documents are then removed from the pool and the process repeated until all acceptable pairs have been collected for that domain.

The sentence alignment approach to identifying aligned documents in (Thompson and Koehn, 2020) uses Vecalign (Thompson and Koehn, 2019) and LASER embeddings to perform sentence alignment and judge sentence similarity. While we use a different aligner and embeddings, our approach follows the same general strategy, with the main difference being that language identification is part of their scoring function while we simply require over half the sentences in each document to obtain an alignment. We can do this as the provided data set we work with has been selected based on

language identification, so we can assume the sentences we work with are generally in the correct language.

Our process results in 4,372 document pairs, containing 160,787 sentence pairs after deduplication. We remove all sentence pairs that have less than three tokens in either language, disregarding all numbers and other non-alphabetical symbols. Furthermore, we remove all sentence pairs that obtain a LaBSE score lower than 0.4. While we do not have any statistics on what the ideal LaBSE threshold should be for this language pair, Steingrímsson et al. (2023a) show that for Icelandic-English over half the sentence pairs are acceptable when the LaBSE score exceeds 0.4, and we base our threshold on that. Our approach results in a set of 120,756 sentence pairs obtained from parallel documents, with 114,301 of those used for training after we have removed sentences that may overlap with test and development datasets.

4.2 Filtering Sentence Pair Candidates

Having extracted sentence pairs from aligned documents, we have yet to consider most of the data in the provided dataset. We experiment with various filtering filtering approaches and as a starting point we simply use the sentence pair candidates provided by the tasks organizers, eight Lithuanian sentences for each Estonian sentence and eight Estonian sentences for each Lithuanian sentence, as described in Section 3. To extract the best sentence pairs, we apply a number of diverse filtering approaches to these sentence pair candidates.

We start by filtering the sets of Estonian and Lithuanian sentences separately:

1. To start with, we have 142,516,521 sentences in Estonian and 210,914,146 sentences in Lithuanian. We deduplicate these sets, giving us 53,228,455 Estonian sentences and 63,536,939 Lithuanian sentences.

2. Although the Fasttext language detection model has been applied to the data, it still contains sentences that are in different languages. In order to remove these we run two additional language detection tools, lingua[5] and langdetect (Shuyo, 2010). From both of these tools we acquire a language classification for each sentence. We then remove all sentences that do not obtain the expected classification by at least two of the three classifiers that have been applied. This leaves us with 33,500,758 Estonian sentences and 43,173,412 lithuanian sentences.

3. In order to remove sentences that may be disfluent we use two pre-trained GPT-2 (Radford et al., 2019) models, one for each language,[6] to classify the sentences. For that we use the approach described in (Steingrímsson et al., 2023a): We collect two sets of sentences for each language, one containing sentences that are presumably fluent and the other one containing sentences that are likely to be disfluent. To train the classifiers, we selected 15,000 sentences randomly for each language from the Leipzig Corpora Collection (Goldhahn et al., 2012) for the fluent examples and 15,000 random sentences from the provided data we had already discarded in the previous step. The classifier uses the GPT-2 model to calculate perplexity for the sentences, and chooses potential thresholds as the average between two adjacent perplexity values. It then uses a maximization function to decide on a threshold that yields the most accurate prediction based on the training set. After classifying the remaining sentences, and removing the approximately 120 thousand sentences included in the document alignment data previously acquired, we are left with 31,298,451 Estonian sentences and 29,498,886 Lithuanian sentences.

Next, we consider the provided sentence pair candidates as calculated using LASER2. We have two candidate lists, one with eight candidates for each Estonian sentence and another with eight candidates for each Lithuanian sentence. We remove all pairs containing sentences not in our filtered sentence lists. We then take an intersection of the resulting sets. The intersection thus only contains sentence pairs where the Lithuanian sentence is one of the top 8 candidates for the Estonian sentence, and vice versa. This gives us a list of 36,250,870 sentence pairs, 35,720,955 after we have excluded all pairs containing sentences that overlap with sentences in the evaluation or development data sets. It should be noted that at this stage some sentences

[5] https://pemistahl.github.io/lingua-py

[6] Lithuanian model: https://huggingface.co/DeividasM/gpt2_lithuanian_small; Estonian model: https://huggingface.co/tartuNLP/gpt-4-est-base

are found in multiple sentence pairs. We proceed to filter this set of sentence pairs:

4. For each Estonian sentence we select only the Lithuanian sentence that gives the highest LASER2 score, and for each Lithuanian sentence we likewise select only the Estonian sentence with the highest score. This reduces the candidate list to 24,735,722 sentence pairs.

5. The sentences comprising the pairs are tokenized. We then run fast-align (Dyer et al., 2013) to obtain word alignments for each sentence pair. These word alignments are used to calculate a word alignment score, WAScore, a word alignment-based score devised to measure word-level parallelism, introduced in Steingrímsson et al. (2021). Steingrímsson et al. (2023a) show that when WAScore is low, very few sentences are good mutual translations. We remove all sentence pairs that have a WAScore lower than 0.15, indicating that 40% or fewer tokens in either sentence obtained an alignment on average. After that our candidate list contains 21,387,140 sentence pairs.

6. We calculate a LaBSE score for all the pairs. If the LaBSE score is higher than 0.9, we accept the sentence pair for our final training set without further processing. These are 891,313 sentence pairs. We also set a minimum threshold of 0.6, as suggested by Feng et al. (2022). This gives us 13,289,869 sentence pairs to processed further, and the rest is discarded.

7. Next, we train Bicleaner AI (Zaragoza-Bernabeu et al., 2022) to classify the Estonian-Lithuanian language pair. For training Bicleaner we need monolingual corpora and parallel corpora. For monolingual data we collected 5 million sentences in each language from the Leipzig Corpora Collection and used 100 thousand parallel pairs randomly selected from the set of sentence pairs extracted from the document alignment step described in Section 4.1. Our Bicleaner AI model gives low scores and we accept sentence pairs with scores over the threshold of 0.05. We run the model on all unfiltered sentences, removing over 20 million and leaving us with 14,988,586 sentence pairs, as shown in Table 1.[7] We later take an intersection of this set and the set obtained by applying other filters, as shown in Table 2.

8. Finally, we use the LASER2 scores, LaBSE scores, WAScore and NMTScore (Vamvas and Sennrich, 2022) with a classifier to predict whether a sentence pair contains a mutual translation. NMTScore is based on translation cross-likelihood, the likelihood that a translation of segment *A* into some language, could also be a translation of segment *B* into the same language. We used OPUS-MT models to translate the segments. We use a logistic regression (Cox, 1958) classifier trained on the same data as the GPT-2 classifiers described above. The classifier accepts as valid mutual translations, 2,967,348 sentence pairs out of the 13,289,869 marked for further processing in (6). When these are added to the set of previously accepted sentences from the aligned documents and the ones having very high LaBSE scores, we have 3,902,740 in our final training set, before applying the Bicleaner AI filter, as shown in Table 2.

5 Results

In addition to the baseline models described in Section 3, we trained eleven MT models using data sets at different stages of the compilation process and evaluated on the provided test sets, using BLEU[8] and chrF[9]. Table 1 shows the results after each filtering step until the logistic regression filter, and Table 2 shows the final sets after filtering and an ablation study on the effects of combining the sets acquired using different approaches. Our best model (**K**) was trained on a combination of sentence pairs from the aligned document pairs (**G**), sentence pairs with a LaBSE score over 0.9 (**H**) and the sentence pairs accepted by our logistic regression filter (**I**).[10]

[7] Our model is available at Github: `https://github.com/steinst/BicleanerAI-models`

[8] Sacrebleu signature: BLEU+nrefs.1+case.mixed+eff.no+tok.3a+smooth.exp+version.2.3.1

[9] Sacrebleu signature: chrF2+nrefs.1+case.mixed+eff.yes+nc.6+nw.0+space:no+version.2.3.1

[10] We submitted dataset *L* to the shared task, which has somewhat lower scores than dataset *K* and was the dataset that was used to train our second best model. This was due to an error in our training script used for selecting a dataset to submit. The script did not remove sentences overlapping with evaluation data, giving us incorrect results. This error has been rectified in all results given in this paper and when we talk about our best model we are always referring to the model trained on dataset *K*.

Data Filters	No. sent.	Bleu				ChrF			
		EMEA	EUB	EP	JRC	EMEA	EUB	EP	JRC
A. Unfiltered	35,720,955	16.2	14.8	15.1	18.2	45.0	43.3	45.9	45.8
B. A ∩ Bicleaner AI	14,988,586	18.7	17.4	17.3	21.8	49.2	48.0	49.5	50.2
C. A ∩ Best LASER2	24,735,722	15.1	15.1	14.7	18.2	45.9	45.3	46.3	48.3
D. C ∩ WAScore filter	21,387,140	19.4	18.9	17.3	23.9	49.3	48.7	49.0	52.0
E. D ∩ LaBSE > 0.6	13,958,582	19.9	19.0	18.3	23.3	50.3	50.6	50.3	52.4
F. B ∩ E	7,193,830	20.5	19.4	18.5	24.2	51.2	51.6	51.4	53.5

Table 1: Scores for the models trained on datasets compiled by applying different filters. We evaluate on the four provided test sets, with data from EMEA, EUBookshop (EUB), Europarl (EP) and JRC-Acquis. The table shows the number of sentences, BLEU and ChrF scores after different filters have been applied.

Data Filters	No. sent.	Bleu				ChrF			
		EMEA	EUB	EP	JRC	EMEA	EUB	EP	JRC
Baseline	2,654,090	18.2	19.1	17.8	24.3	49.5	52.2	51.5	54.8
G. Aligned Docs	114,301	8.0	10.9	9.3	16.2	33.8	41.6	40.3	44.5
H. LaBSE > 0.9	868,039	18.9	17.2	16.3	22.6	50.1	50.3	49.9	52.6
I. Logistic Regression	2,925,549	15.4	14.1	13.7	18.2	45.7	46.2	46.5	48.0
J. H ∪ I	3,788,511	20.2	19.5	18.3	24.8	51.2	52.1	51.7	54.4
K. G ∪ H ∪ I	3,902,740	20.4	20.7	19.1	26.6	51.4	53.3	52.2	56.1
L. K ∩ B	2,684,931	20.4	19.7	18.4	25.1	51.4	52.5	51.8	54.9

Table 2: Datasets created using different approaches and an ablation study for investigating the effect of each dataset on MT quality as measured by BLEU and ChrF. The logistic regression dataset is created by applying our logistic regression classifier on dataset E in Table 1. We evaluate on the four provided test sets.

Our best model outperforms the baseline by approximately 1.9 BLEU on average. We find that the sentence pairs from the aligned documents, only 114,301 pairs, improve the BLEU on average by 1.0 BLEU. This indicates that these sentence pairs are useful and that identifying document alignments in web-scraped data is worth the effort. We also find that the sentence pairs having high LaBSE scores, over 0.9, give much better results on their own than over three times more sentence pairs with LaBSE scores in the range 0.6 to 0.9, even though they have been filtered further using additional methods. As shown in Table 2, combining these two sets raises the scores substantially. Furthermore, while the Bicleaner AI model we trained seemed to give decent results in earlier stages, using it to filter the dataset we acquired using other approaches actually decreased the scores. This indicates that the Bicleaner AI model is rejecting too many useful sentence pairs. It could be useful to try to investigate further which of these rejected sentences are useful for MT training and which are truly detrimental, but we leave that for future work.

6 Conclusions and Future Work

Our alignment and filtering approach resulted in an improvement over the baseline in terms of BLEU score for the four evaluation sets. We identified 4,372 document pairs in the provided dataset, which we aligned on sentence level and used the resulting data set for training. We then combined a number of filtering approaches for determining which sentence pair candidates from a provided candidate list would be likely to be useful, these included an ensemble approach for language detection, using three different tools, a GPT-2 based classifier to determine whether sentences are fluent or disfluent, a logistic regression classifier based on word alignment scores and two sentence embedding based scores, LaBSE and LASER2, and finally a Bicleaner AI classifier.

Working in a similar vein, many different paths could be taken for future work on this problem. Steingrímsson et al. (2023a) show that it can be beneficial to inspect how different filters suit a given translation direction. A filtering method giving an optimal results for $\text{lang}_a \rightarrow \text{lang}_b$ is not necessarily the optimal filtering approach for $\text{lang}_b \rightarrow \text{lang}_a$.

In this work we did not try to evaluate the filtering approaches with regards to translation direction. For translating only from Estonian to Lithuanian, removing incoherent and ungrammatical Estonian sentences may not be as important as removing such sentences in Lithuanian, as it is more important that the target language data contains coherent and well written examples. Different levels of filtering for the different languages could thus be useful in order to add more useful examples.

The aim of our filters is to remove sentences likely to be detrimental in MT training. While we do know about some of the qualities that reduce translation quality, as discussed in Section 2, more fine-grained classifications may be useful. For example, we could designate different levels of misalignments, which include partial alignments defined as sentence pairs where a part of one or both sentences is not represented in the other sentence. Steingrímsson et al. (2023c) argue that extracting mutual translations from such pairs, while discarding the extraneous data, may improve the quality of MT models trained on the data, and show that for one parallel corpus. If that holds in general, it could be useful when working with web-scraped data to identify when misalignments become detrimental and when they can be useful, as well as helping to come up with effective ways to refine such sentence pairs.

Table 2 shows that the datasets compiled from the aligned documents and the one comprising sentence pairs with very high LaBSE scores are very useful as additional training data. We presume that this is an indication of these sets containing higher-quality data. While not suitable for the shared task, it would be an interesting experiment to use a curriculum learning approach for training models on web-scraped corpora such as the one we are using by training a model first on a large set of possibly useful sentences and then fine-tuning the model on the higher-quality data.

Finally, it should be noted that the training times for these models varied considerably. While our best model reached the optimal checkpoint in approximately 20 hours and the second best in 12 hours, the models trained on the larger datasets listed in Table 1 took between 50 and 80 hours of training, using the same settings, while still resulting in lower quality models. It shows that careful curation of training data for MT is not only important for improving model quality in terms of better translations, it also allows for much faster training resulting in a lower carbon footprint.

References

Mikel Artetxe and Holger Schwenk. 2019a. Margin-based Parallel Corpus Mining with Multilingual Sentence Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy.

Mikel Artetxe and Holger Schwenk. 2019b. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Mikko Aulamo, Sami Virpioja, and Jörg Tiedemann. 2020. OpusFilter: A Configurable Parallel Corpus Filtering Toolbox. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 150–156, Online.

Christian Buck and Philipp Koehn. 2016. Findings of the WMT 2016 bilingual document alignment shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 554–563, Berlin, Germany.

Vishrav Chaudhary, Yuqing Tang, Francisco Guzmán, Holger Schwenk, and Philipp Koehn. 2019. Low-Resource Corpus Filtering Using Multilingual Sentence Embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 261–266, Florence, Italy.

David R Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232.

Aswarth Abhilash Dara and Yiu-Chang Lin. 2016. YODA system for WMT16 shared task: Bilingual document alignment. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 679–684, Berlin, Germany.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT Sentence Embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland.

William A. Gale and Kenneth W. Church. 1991. A Program for Aligning Sentences in Bilingual Corpora. In *Proceedings of the 29th Annual Meeting on Association for Computational Linguistics*, Berkeley, California.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey.

Luís Gomes and Gabriel Pereira Lopes. 2016. First steps towards coverage-based document alignment. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 697–702, Berlin, Germany.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. *CoRR*, abs/2205.12654.

Christian Herold, Jan Rosendahl, Joris Vanvinckenroye, and Hermann Ney. 2022. Detecting Various Types of Noise for Neural Machine Translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2542–2551, Dublin, Ireland.

Felix Hieber, Michael Denkowski, Tobias Domhan, Barbara Darques Barros, Celina Dong Ye, Xing Niu, Cuong Hoang, Ke Tran, Benjamin Hsu, Maria Nadejde, Surafel Lakew, Prashant Mathur, Anna Currey, and Marcello Federico. 2022. Sockeye 3: Fast neural machine translation with pytorch.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain.

Huda Khayrallah and Philipp Koehn. 2018. On the Impact of Various Types of Noise on Neural Machine Translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia.

Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the WMT 2020 Shared Task on Parallel Corpus Filtering and Alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742, Online.

Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino. 2019. Findings of the WMT 2019 Shared Task on Parallel Corpus Filtering for Low-Resource Conditions. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 54–72, Florence, Italy.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the WMT 2018 Shared Task on Parallel Corpus Filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 726–739, Belgium, Brussels.

Sainik Mahata, Dipankar Das, and Santanu Pal. 2016. WMT2016: A hybrid approach to bilingual document alignment. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 724–727, Berlin, Germany.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz Rojas. 2020. Bifixer and Bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal.

Nick Rossenbach, Jan Rosendahl, Yunsu Kim, Miguel Graça, Aman Gokrani, and Hermann Ney. 2018. The RWTH Aachen University Filtering System for the WMT 2018 Parallel Corpus Filtering Task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 946–954, Belgium, Brussels.

Víctor M. Sánchez-Cartagena, Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez. 2018. Prompsit's submission to WMT 2018 Parallel Corpus Filtering shared task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 955–962, Belgium, Brussels.

Rico Sennrich and Martin Volk. 2010. MT-based Sentence Alignment for OCR-generated Parallel Texts. In *The Ninth Conference of the Association for Machine Translation in the Americas*, Denver, Colorado.

Nakatani Shuyo. 2010. Language detection library for java.

Steve Sloto, Brian Thompson, Huda Khayrallah, Tobias Domhan, Thamme Gowda, and Philipp Koehn. 2023. Findings of the WMT 2023 shared task on parallel data curation. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore. Association for Computational Linguistics.

Steinþór Steingrímsson, Pintu Lohar, Hrafn Loftsson, and Andy Way. 2021. Effective Bitext Extraction From Comparable Corpora Using a Combination of Three Different Approaches. In *Proceedings of the*

14th Workshop on Building and Using Comparable Corpora (BUCC 2021), pages 8–17, Online (Virtual Mode).

Steinþór Steingrímsson. 2023. *Effectively compiling parallel corpora for machine translation in resource-scarce conditions*. Ph.D. thesis, Reykjavik University.

Steinþór Steingrímsson, Hrafn Loftsson, and Andy Way. 2023a. Filtering matters: Experiments in filtering training sets for machine translation. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 588–600, Tórshavn, Faroe Islands.

Steinþór Steingrímsson, Hrafn Loftsson, and Andy Way. 2023b. Sentalign: Accurate and scalable sentence alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Singapore, Singapore. Association for Computational Linguistics.

Steinþór Steingrímsson, Pintu Lohar, Hrafn Loftsson, and Andy Way. 2023c. Do not discard – extracting useful fragments from low-quality parallel data to improve machine translation. In *Proceedings of the Second Workshop on Corpus Generation and Corpus Augmentation for Machine Translation*, pages 1–13, Macau, China.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved Sentence Alignment in Linear Time and Space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China.

Brian Thompson and Philipp Koehn. 2020. Exploiting sentence order in document alignment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5997–6007, Online. Association for Computational Linguistics.

Jannis Vamvas and Rico Sennrich. 2022. NMTScore: A multilingual analysis of translation-based text similarity measures. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 198–213, Abu Dhabi, United Arab Emirates.

Jaume Zaragoza-Bernabeu, Gema Ramírez-Sánchez, Marta Bañón, and Sergio Ortiz Rojas. 2022. Bicleaner AI: Bicleaner Goes Neural. In *Proceedings of the Language Resources and Evaluation Conference*, pages 824–831, Marseille, France.

Document-Level Language Models for Machine Translation

Frithjof Petrick[1,2] **Christian Herold**[1,2] **Pavel Petrushkov**[1]
Shahram Khadivi[1] **Hermann Ney**[2]
[1]eBay, Inc., Aachen, Germany
{cherold, ppetrushkov, skhadivi}@ebay.com
[2]Human Language Technology and Pattern Recognition Group
RWTH Aachen University, Aachen, Germany
{petrick, ney}@i6.informatik.rwth-aachen.de

Abstract

Despite the known limitations, most machine translation systems today still operate on the sentence-level. One reason for this is, that most parallel training data is only sentence-level aligned, without document-level meta information available. In this work, we set out to build context-aware translation systems utilizing document-level monolingual data instead. This can be achieved by combining any existing sentence-level translation model with a document-level language model. We improve existing approaches by leveraging recent advancements in model combination. Additionally, we propose novel weighting techniques that make the system combination more flexible and significantly reduce computational overhead. In a comprehensive evaluation on four diverse translation tasks, we show that our extensions improve document-targeted scores substantially and are also computationally more efficient. However, we also find that in most scenarios, back-translation gives even better results, at the cost of having to re-train the translation system. Finally, we explore language model fusion in the light of recent advancements in large language models. Our findings suggest that there might be strong potential in utilizing large language models via model combination.

1 Introduction

Machine translation (MT), the automatic translation of text from one language to another, has seen significant advancements in recent years, primarily driven by neural machine translation (NMT) models (Bahdanau et al., 2015; Vaswani et al., 2017). These models have demonstrated remarkable capabilities in capturing complex linguistic patterns and producing high-quality translations (Wu et al., 2016; Hassan et al., 2018). Nevertheless, most models to-date operate on sentence-level, i.e. translate sentences independently without the context of the surrounding document. Without access to such context, it is impossible for these MT systems to account for discourse-level phenomena such as resolution of ambiguous words and coherence. Unsurprisingly, automatic translations are perceived as much worse, when they are evaluated on entire documents rather than just at the sentence-level (Läubli et al., 2018, 2020; Maruf et al., 2022).

An obvious solution to this problem is to utilize context-aware MT models (Tiedemann and Scherrer, 2017). While document-level NMT models have been thoroughly studied in recent years, sentence-level MT remains the standard despite its inherent limitations. One of the main reasons for this is that most of the document-level approaches rely on parallel training data with document-level metadata. Most releases of large parallel training corpora lack this information and remain purely sentence-level (Bañón et al., 2020; Schwenk et al., 2021). In contrast, large amounts of document-level monolingual data are readily available for almost all domains and languages.

In this work, we strive to build a context-aware MT system that does not rely on any parallel document-level training data. Instead, we use monolingual documents to train a document-level language model (LM), which we fuse with an existing sentence-level MT model during translation. While existing work on LM fusion shows that the fused model is able to incorporate document-level context (Jean and Cho, 2020; Sugiyama and Yoshinaga, 2021), these approaches can be improved. Our work aims to do so in two main directions.

First, we acknowledge that NMT models implicitly learn the language modeling task during training. Recently, Herold et al. (2023) showed that estimating and neutralizing this internal LM can improve translation quality for sentence-level MT. We adapt their approach to document-level LM fusion and demonstrate that this also improves discourse modeling.

Second, the contribution of the fused MT model,

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 375–391
December 6–7, 2023.

the document-level LM and the internal LM must be balanced by a set of fusion scales. Existing work defines the fusion scales as static hyperparameters which are tuned on a validation set via an extensive grid search (Gülçehre et al., 2015; Jean and Cho, 2020; Sugiyama and Yoshinaga, 2021). In our work, we provide two simple alternatives to grid search which allow for automatically tuned context-dependent fusion scales. Our approaches eliminate the need for expensive tuning and further improve discourse-modelling.

The contributions of this work are as follows:

1. We propose multiple extensions to the existing approaches on document-level LM fusion for MT.

2. We compare our methods against two strong baselines: Back-translation, the to-date most popular way to utilize monolingual data for MT, and a task-specific LM re-ranking baseline for pronoun disambiguation. The comparison takes place over four diverse translation tasks in terms of general translation quality as well as specific context-dependant phenomena.

3. We present first results on fusing a large language model (LLM) with a sentence-level MT system.

2 Related Works

Most works on document-level NMT rely on parallel document-level data for system training.

Tiedemann and Scherrer (2017) propose to concatenate adjacent sentences on source and target side and input this into the NMT model which has the exact same architecture as the vanilla sentence-level transformer (Vaswani et al., 2017). Later, many works have proposed modifications to the architecture to better accommodate the additional context (Jean et al., 2017; Bawden et al., 2018; Zhang et al., 2018; Voita et al., 2018; Kuang and Xiong, 2018; Miculicich et al., 2018; Maruf and Haffari, 2018). However, it has been shown that the simple concatenation approach performs as good, if not better than these more complicated variants (Lopes et al., 2020; Sun et al., 2022).

Maybe the biggest challenge for document-level NMT is that most of the parallel MT training data is not document-level (Esplà-Gomis et al., 2019; Schwenk et al., 2021). Recently there has been some effort to restore document-level meta information from existing sentence-level corpora but this is a very time consuming and error-prone process (Ghussin et al., 2023). Therefore, approaches to document-level NMT have been proposed that utilize document-level monolingual data, of which typically large amounts are readily available.

One direction is to back-translate the document-level monolingual data to create synthetic parallel document-level data. The reverse system used for back-translation can be either sentence-level (Junczys-Dowmunt, 2019; Saleh et al., 2019; Post and Junczys-Dowmunt, 2023) or document-level (Sugiyama and Yoshinaga, 2019; Huo et al., 2020). A downside of this approach is that the final MT system has to be re-trained to incorporate the new synthetic data.

Another line of work uses document-level language models in combination with sentence-level translation models. Gülçehre et al. (2015) were the first to propose a log-linear combination of sentence-level language and NMT models, coining the term 'shallow fusion'. Recently, it was shown that the shallow fusion approach for sentence-level NMT can be improved by compensating for the implicitly learned internal language model of the NMT system (Herold et al., 2023). Regarding the integration of a document-level LM, earlier approaches simply use the LM for re-ranking the hypothesis of the sentence-level NMT model (Stahlberg et al., 2019; Yu et al., 2020). Several works have proposed to employ a log-linear combination between sentence-level NMT system and document-level LM (Garcia et al., 2019; Jean and Cho, 2020; Sugiyama and Yoshinaga, 2020). Both Jean and Cho (2020) and Sugiyama and Yoshinaga (2020) propose to also include the probabilities of the LM without context information in order to mitigate the influence of the current sentence on the LM probabilities. While our approach also uses the output of a sentence-level LM, it is conceptually different from the previous works in that we want to mitigate the influence of the internal LM from the NMT model, resulting in a different final formulation. To further improve LM incorporation, Jean and Cho (2020) propose to use subword-dependent fusion scales instead of a single scale per model.

Apart from back-translation and LM integration there exist some other ways to utilize additional monolingual document-level data for MT. Voita et al. (2019) train a document-level automatic post

editing system on the monolingual data and use it to improve the hypotheses from a sentence-level NMT system in a two-pass approach. Several works utilize the additional data in a multi-task learning approach (Junczys-Dowmunt, 2019) or for pre-training (Zhu et al., 2020; Chen et al., 2021b; Liu et al., 2020; Chen et al., 2021a).

Very recently, LLMs have shown their potential for the task of document-level NMT (Wang et al., 2023). However, it is unclear how much parallel training samples were seen during the large scale pre-training on trillions of tokens.

3 Document-level Language Model Fusion

The sentence-level MT model translates a source sentence F into a target sentence $E := e_0^I$ of subwords e_i. In the document-level LM fusion approach, we additionally provide the k previous target-side sentences E_{-k}^{-1} as context[1].

3.1 Internal Language Model Neutralization

As the translation model already implicitly learns probabilities that are source-independent, directly fusing the MT model and the document-level LM overvalues the source-agnostic probabilities. Therefore, we estimate the internal LM of the MT model and in total combine three models during generation:

- the existing sentence-level MT model $p_{\text{TM}}(e_i) := p_{\text{TM}}(e_i \,|\, e_0^{i-1}, F)$,
- the LM $p_{\text{LM}}(e_i) := p_{\text{LM}}(e_i \mid e_0^{i-1}, E_{-k}^{-1})$ trained on monolingual documents with access to the previous target sentences E_{-k}^{-1},
- and a second LM $p_{\text{ILM}}(e_i) := p_{\text{ILM}}(e_i \,|\, e_0^{i-1})$ which estimates the internal LM probabilities implicitly learned by the MT model. We train this LM separately on the target-side of the MT training data, as we found that this approach works best for document-level MT when compared to other approaches presented by Herold et al. (2023). This comparison can be found in Appendix A.3.

We multiply the model output probabilities and normalize them. The resulting probability distribution is now conditioned on both the source sentence F and the target-side context E_{-k}^{-1}:

$$p(e_i) := p(e_i \,|\, e_0^{i-1}, F, E_{-k}^{-1}) := \frac{p_{\text{TM}}^{\lambda_0}(e_i) \cdot p_{\text{LM}}^{\lambda_1}(e_i) \cdot p_{\text{ILM}}^{-\lambda_2}(e_i)}{\sum_{e'} p_{\text{TM}}^{\lambda_0}(e') \cdot p_{\text{LM}}^{\lambda_1}(e') \cdot p_{\text{ILM}}^{-\lambda_2}(e')}. \quad (1)$$

Each model is weighted with a scalar $\lambda_0, \lambda_1, \lambda_2 \geq 0$, the internal LM is included with a negative exponent. We tune these fusion scales on the validation set for BLEU via a grid search over $\lambda_0, \lambda_1, \lambda_2 \in \{0, 0.1, \ldots, 1\}$.

Existing work on document-level LM fusion uses a similar formulation as our approach, but instead of neutralizing the internal LM of the MT model, it accounts for the sentence-level probabilities $p_{\text{LM}}(e_i \,|\, e_0^{i-1})$ of the document-level LM (Jean and Cho, 2020; Sugiyama and Yoshinaga, 2021). In the particular case where there are no previous sentences available, this approach simply falls back to using only the sentence-level MT model probabilities. Our approach on the contrary can also leverage the gains obtained from sentence-level LM fusion and is theoretically more expressive.

3.2 Context-dependent Fusion Scales

Choosing appropriate fusion scales $\lambda_0, \lambda_1, \lambda_2$ in Equation 1 is crucial. Conventionally, the scales are tuned via grid search. This is problematic in three aspects:

1. Grid search is expensive. Testing e.g. ten possible values for each of the three model scales already requires translating the validation set 1000 times.
2. The tuning process depends on the tuning data, its domain and the tuning objective. E.g., the scales that optimize document-targeted metrics differ from the ones that maximize sentence-level translation quality (Sugiyama and Yoshinaga, 2021).
3. Fusion scales obtained by a hyperparameter grid search must be constant. Document-level context however is not uniformly useful for all predicted subwords.

In the following, we propose two simple alternatives to obtaining fusion scales with grid search that overcome the aforementioned issues.

[1] At the beginning of the document we only provide as many sentences as available.

3.2.1 On-the-fly Fusion Scales

During decoding, the next subword e_i is chosen to maximize the fused probability (Equation 1). We propose to also choose the fusion scales in a similar fashion and define them to maximize the fused model scores:

$$(\lambda_0, \lambda_1, \lambda_2) \coloneqq \underset{(\lambda_0,\lambda_1,\lambda_2)}{\operatorname{argmax}} \frac{p_{\text{TM}}^{\lambda_0}(e_i) \cdot p_{\text{LM}}^{\lambda_1}(e_i) \cdot p_{\text{ILM}}^{-\lambda_2}(e_i)}{\sum_{e'} p_{\text{TM}}^{\lambda_0}(e') \cdot p_{\text{LM}}^{\lambda_1}(e') \cdot p_{\text{ILM}}^{-\lambda_2}(e')}. \quad (2)$$

Our model maximizes over the discrete set $\lambda_0, \lambda_1, \lambda_2 \in \{0, 0.1, \ldots, 1\}$. This approach obviates the need for separate scale tuning entirely and only has a small overhead during generation.

3.2.2 Automatically Learned Fusion Scales

Alternatively, we propose to learn the fusion scales automatically using a small amount of training examples (F, E, E_{-k}^{-1}) with document-level context, similarly to Jean and Cho (2020). We obtain the training data by back-translating the monolingual data (see Section 5). Automatic learning allows us to implement subword-dependent fusion scales: We introduce a set of learnable parameters $\lambda_0(e), \lambda_1(e), \lambda_2(e)$ for each subword e from the target vocabulary and learn them automatically by optimizing the cross-entropy loss

$$(\lambda_0, \lambda_1, \lambda_2) \coloneqq \underset{\lambda\colon V \to \mathbb{R}^3}{\operatorname{argmax}} \sum_{(F,E,E_{-k}^{-1})} \sum_i \log \frac{p_{\text{TM}}^{\lambda_0(e_i)}(e_i) \cdot p_{\text{LM}}^{\lambda_1(e_i)}(e_i) \cdot p_{\text{ILM}}^{-\lambda_2(e_i)}(e_i)}{\sum_{e'} p_{\text{TM}}^{\lambda_0(e')}(e') \cdot p_{\text{LM}}^{\lambda_1(e')}(e') \cdot p_{\text{ILM}}^{-\lambda_2(e')}(e')}. \quad (3)$$

Scale learning uses the same optimization parameters as the MT model was originally trained with. The scale parameters are initialized with a small variance around zero while all other parameters are frozen.

4 Document-level Language Model Pronoun Re-ranking

Besides consistency, the main problem of discourse-modelling are ambiguities. E.g. translating the English pronoun 'it' to German requires access to the noun that it refers to, which might only be found in a preceding sentence (Müller et al., 2019).

We propose an approach specific to the En→De language pair that directly targets the pronoun translation problem by re-ranking sentence-level hypotheses using a document-level LM. We first translate each sentence independently using the sentence-level MT model. Each sentence-level translation is expanded to a set of candidates by replacing the pronouns with all alternatives ('er', 'sie', 'es'). All candidate translations are then scored in context of the preceding sentences using a document-level LM, and we select the pronoun for which the LM score is highest.

This approach is very much tailored to the specific pronoun translation problem for this specific language pair. While it is theoretically possible to extend this approach to cover more cases, this will require extensive human effort and is probably not feasible in most scenarios. However, we include it here, because it serves as a reasonable baseline for this popular pronoun translation benchmark.

5 Document-level Back-translation

The to-date most popular way of utilizing monolingual data for MT is to create synthetic parallel training data via back-translation (Sennrich et al., 2016). We train a sentence-level backwards MT system on the parallel data and use it to translate the document-level monolingual data back into the source language. The sentence-level translations are concatenated to obtain synthetic parallel documents (Junczys-Dowmunt, 2019; Saleh et al., 2019; Sugiyama and Yoshinaga, 2019; Huo et al., 2020; Post and Junczys-Dowmunt, 2023).

To train the final systems we combine the authentic sentence-level parallel and the synthetic document-level data. Combining both data sources is not straightforward, because of their varying size and the difference between sentence/document-level context. Therefore, we first oversample the data accordingly to have roughly the same number of sentences in both parts. Secondly, we turn the authentic sentence-level parallel data into 'pseudo-documents' by concatenating them in a random order (Junczys-Dowmunt, 2019; Jean et al., 2019). This ensures that all training data has the same context size. We found this procedure to perform best when incorporating synthetic document-level data. For a detailed comparison, see Appendix A.5.

6 Experiments

6.1 Tasks

We evaluate our approaches on four different tasks of varying data conditions and domains. Three

tasks are on publicly available data and a fourth task is based on a large scale internal dataset in the e-Commerce domain. All tasks include (sentence-level) parallel training data and document-level monolingual data from the same domain. The exact data conditions are provided in Appendix A.1.

The *News En→De* data consists of news articles while the *TED En→It* task consists of scientific talks. Both are low resource with less than 1M training samples in total. The *Subtitles En→De* data consists of subtitles from various TV shows and is medium size. Finally, the *e-Commerce En→De* task is about translating item descriptions from e-Commerce listings and the training data is large scale with more than 100M examples.

While the parallel training data for the three academic tasks does provide document-level metadata, our approaches do not make use of this information and we assume that the parallel training data is sentence-level for most experiments. We only make use of this information to provide a direct comparison against the setting where document-level parallel data is assumed to be available. As ParaCrawl, like most other large-scale web-crawled parallel datasets, is not a document-level corpus, we can not conduct these experiments for the e-Commerce task.

We preprocess each corpus with byte-pair encodings (Sennrich et al., 2016) using the SentencePiece toolkit (Kudo, 2018) learned on the parallel dataset with a shared vocabulary of 32k subwords (13.6k for TED). For the e-Commerce task we additionally use inline casing (Berard et al., 2019; Etchegoyhen and Gete, 2020).

6.2 Settings

We train transformer MT models in the 'base' configuration (Vaswani et al., 2017), implemented in Fairseq (Ott et al., 2019). For the LMs we use a similar architecture but without the encoder. Our document-level models use the same architecture as the sentence-level models, we simply include context sentences by concatenating the previous two source and target sentences to all training examples, separated by a reserved symbol (Tiedemann and Scherrer, 2017).

Details on the optimization algorithm are given in Appendix A.1. The final model is selected based on the validation set perplexity. We then perform beam search with beam size 12 and length normalization. Document-level decoding uses the 'last sentence' search strategy as described in Herold and Ney (2023b).

The document-level LMs are trained on a combination of target-side of the sentence-level parallel and document-level monolingual data. Regardless of the task, we train the LMs for 300k update steps with batch size 90k, 10 % dropout, and 10 % label smoothing.

For the LM fusion experiments with non-static fusion scales, we restrict the search space to only consider scale combinations where $\lambda_0 = 1$ and $\lambda_1 = \lambda_2$. A direction comparison is given in Appendix A.4. For back-translation, we use beam search with beam size 4 and increase the training time proportionally to the new data size.

6.3 Evaluation

Document-level evaluation is challenging, as intersentential context usually is only relevant for a small fraction of words. Further, conventional metrics like BLEU (Papineni et al., 2002) or COMET (Rei et al., 2020) do not appropriately measure how well document-level context is considered for those words where context does matter (Läubli et al., 2018, 2020; Maruf et al., 2022). However, we still report BLEU using Sacrebleu (Post, 2018) and COMET[2] on the task-specific in-domain test sets to evaluate the general MT quality.

To better evaluate the improvements from the document-level approaches, we focus on selected sentences for which document-level context is known to be important. Here, we report on two test sets focusing on ambiguities. The En→De *pronouns* test set released by Müller et al. (2018) was curated from OpenSubtitles shows and contains 12k examples. Most examples require previous sentences as context to properly translate the English pronoun 'it' with German 'er', 'sie' or 'es'. Further, the *gender-referring professions* test sets released as contextual part of MT-GenEval (Currey et al., 2022) are available for various target languages and focus on a wider range of ambiguous words, e.g. whether 'the teacher' should be translated with 'die Lehrerin' or 'der Lehrer' in German. Again, context from the previous sentences is required to determine the correct translation. We use these test sets for En→De and En→It which both comprise approx. 1.1k examples that were created by translating Wikipedia articles.

Computing BLEU and COMET on these chal-

[2] Using the `wmt22-comet-da` model (Rei et al., 2020)

lenge test sets better reflects how well a MT system handles document-level context. An even more specific metric can be obtained by focusing only on the ambiguous words. Previous work commonly reports an accuracy metric that is based on contrastive scoring, which is computed by comparing the model probabilities of the reference against a set of contrastive examples (Müller et al., 2018). This metric however can be misleading, as it not based on the generated translation but rather just on scoring. MT systems with high contrastive scores often perform poorly when their generated hypothesis is evaluated (Post and Junczys-Dowmunt, 2023). Instead, we focus on translation-based document-targeted metrics.

On the pronouns test set, we compute a pronoun F1-score as proposed by Herold and Ney (2023a). This metric directly compares the pronouns of the hypothesis and the reference and is based on the BLONDE metric (Jiang et al., 2022). On the professions test set, we report the translation-based accuracy metric suggested by their curators (Currey et al., 2022). Further, for the Subtitles system we also report a formality F1-score on its test set as proposed by Herold and Ney (2023a).

6.4 Results

We evaluate our approaches to utilize monolingual document-level data on the four MT tasks. We apply them in two settings where a) we assume that all parallel data is purely sentence-level, and b) also the parallel data is document-level.

In an effort to compare to previous work, we re-implement LM fusion with static scales without subtracting the internal LM which was independently proposed by Jean and Cho (2020) and Sugiyama and Yoshinaga (2021). These works subtract the intersentential probabilities of the external LM instead. Further, we also re-implement the non-static scales predicted with a 'merging module' learned on parallel document-level data as proposed by Jean and Cho (2020).

We first evaluate our approaches on conventional metrics to measure their general MT performance. Then, we focus on the document-targeted challenge sets to quantify how well they utilize document-level context.

6.4.1 Conventional Metrics

We start by evaluating on the in-domain test sets of the four MT tasks using the conventional MT metrics. Here, we do not expect to see much improvements coming from the document-level context. The results are presented in Table 1.

Adding monolingual data gives the largest improvements on News and small improvements on the e-Commerce task. On these two tasks, the monolingual data is in-domain and the improvements are likely because of the domain. On Subtitles and TED we do not see any improvements as Subtitles already has a large amount of in-domain parallel data and the TED monolingual data is slightly out-of-domain. We verified the domain effect by training sentence-level LMs on equal amounts of data from the target-side of the parallel and monolingual corpora and comparing their perplexities on the test sets. Details are provided in Appendix A.2.

None of the presented approaches significantly decreases translation performance in terms of conventional metrics. The only exception is the back-translation which when added to the Subtitles and TED document-level baseline performs worse in BLEU. In COMET however, this decrease is less prevalent.

6.4.2 Document-targeted Metrics

The results on the document-targeted test sets are shown in Table 2. First we discuss the scenario without access to document-level parallel training data.

LM fusion. Adding monolingual documents to the sentence-level baseline with the existing approaches from Jean and Cho (2020) and Sugiyama and Yoshinaga (2021) improves scores only marginally by on average +0.5 % absolute F1 score on the pronouns test set and no improvements on the professions set. In comparison, our approach on LM fusion with the neutralization of the internal LM performs better: E.g., the variant with on-the-fly scales on average improves the pronoun F1 score by +2.4 % and the professions ac-

[3] External baseline by Herold and Ney (2023b)

[4] External baseline by Huo et al. (2020)

[5] Re-implementation of LM fusion with neutralization of the intersentential LM probabilities instead of the internal LM, as introduced by Jean and Cho (2020) and Sugiyama and Yoshinaga (2021)

[6] Re-implementation of the 'merging module' approach by Jean and Cho (2020). This approach uses parallel document-level data for scale learning.

Data		Method	News		Subtitles		TED		e-Commerce	
parallel	mono.		BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
sent.	-	baseline (prev. work)	32.8[3]	-	37.3[4]	-	34.2[3]	-	-	-
		baseline (ours)	32.7	82.8	37.3	87.9	34.8	86.1	36.4	89.2
	doc.	(Jean, 2020; Sugiyama, 2021)[5]	33.1	83.2	37.2	87.8	34.6	86.2	37.1	89.6
		(Jean, 2020)[6]	32.9	83.0	37.3	87.9	34.5	86.2	36.6	89.2
		LM: static	34.8	84.2	37.2	87.8	34.9	86.2	**37.3**	89.6
		LM: on-the-fly	34.7	83.9	37.2	87.9	34.9	86.2	36.8	**89.7**
		LM: auto. learned	34.4	83.8	37.4	87.8	34.7	86.2	36.8	89.0
		LM: re-rank pronouns	32.6	82.7	36.9	87.8	n.a.		36.4	89.2
		back-translation	37.1	85.2	37.2	87.6	35.1	86.6	36.2	89.3
		+ LM: static	**37.4**	**85.6**	**37.6**	87.7	35.2	86.6	35.0	88.9
		+ LM: on-the-fly	37.2	85.4	37.1	87.6	34.8	86.6	35.9	89.4
		+ LM: auto. learned	37.2	85.3	37.3	87.6	34.9	86.6	36.2	89.5
doc.	-	baseline	32.5	82.9	39.5	**88.2**	**35.4**	86.5	n.a.	
	doc.	LM: static	35.1	84.3	38.9	**88.2**	35.2	**86.7**		
		LM: on-the-fly	34.5	84.1	39.0	88.0	35.1	**86.7**		
		LM: auto. learned	34.8	84.1	39.3	88.2	35.2	86.6		
		LM: re-rank pronouns	32.3	82.8	39.1	88.1	n.a.			
		back-translation	37.2	85.3	37.5	87.8	34.6	86.5		

Table 1: Utilizing document-level monolingual data using different methods, reporting on the in-domain test sets of each task. BLEU and COMET are given in percentage. Best results for each column are highlighted.

Data		Method	News		Subtitles			TED	e-Commerce	
parallel	mono.		pron.	proff.	pron.	proff.	form.	proff.	pron.	proff.
sent.	-	baseline (prev. work)	45.3[3]	-	41.1[3]	-	59.4[3]	-	-	-
		baseline (ours)	45.1	65.9	41.7	65.3	57.2	65.4	42.6	63.7
	doc.	(Jean, 2020; Sugiyama, 2021)[5]	46.0	65.0	42.3	65.8	58.1	65.1	42.7	64.0
		(Jean, 2020)[6]	45.1	64.7	41.9	65.8	57.7	65.4	42.5	63.5
		LM: static	45.5	65.5	42.5	66.3	58.4	65.4	42.8	64.4
		LM: on-the-fly	48.0	65.5	44.2	65.9	58.9	66.4	44.4	66.2
		LM: auto. learned	46.7	64.9	42.8	65.5	58.6	65.6	44.0	65.2
		LM: re-rank pronouns	48.0	66.1	57.5	65.5	57.2	n.a.	**54.5**	64.0
		back-translation	48.7	80.5	52.3	67.0	58.5	65.1	42.9	67.1
		+ LM: static	48.5	80.6	53.1	68.3	53.8	65.4	42.6	66.0
		+ LM: on-the-fly	48.9	**81.3**	52.8	67.3	60.4	65.4	46.3	**70.5**
		+ LM: auto. learned	48.9	80.5	52.0	67.6	59.9	65.4	46.2	65.7
doc.	-	baseline	**55.9**	71.2	67.2	70.8	61.9	67.2	n.a.	
	doc.	LM: static	55.3	70.8	67.5	71.1	61.5	66.8		
		LM: on-the-fly	55.8	72.3	67.8	**71.9**	61.4	**67.6**		
		LM: auto. learned	55.7	71.5	**67.4**	71.0	61.6	**67.6**		
		LM: re-rank pronouns	50.9	71.5	62.6	70.8	61.9	n.a.		
		back-translation	52.1	79.4	62.8	67.3	**62.0**	65.7		

Table 2: Document-targeted evaluation of the different approaches utilizing document-level monolingual data. We report the pronoun F1 score (Herold and Ney, 2023a), gender-referring professions accuracy (Currey et al., 2022) and the formality F1 score on the Subtitles test set (Herold and Ney, 2023a), all given in percentage. Best results for each column are highlighted.

curacy by +0.9 %. Compared to static scales, both on-the-fly and automatically learned scales yield small improvements and further do not involve the expensive grid search.

LM re-ranking pronouns. Our LM re-ranking approach was specifically tailored towards the pronouns test set. We see most improvements on this test set, while the document-targeted metrics on the other test sets remain mostly unchanged. For both the Subtitles and the e-Commerce task, LM re-ranking is the best approach of utilizing document-level monolingual data for this specific test set in the absence of document-level parallel data. On News however, the gains are less prevalent: Our analysis finds that even though the LM in this case can predict the pronouns correctly, the general translation quality of the baseline on this test set is low and therefore this model often fails to generate any pronouns at all. This again highlights the discrepancy between scoring- and generation-based metrics.

Back-translation. In a direct comparison to LM fusion, back-translation outperforms LM fusion despite our improvements over the existing work. Back-translation on average improves the pronouns F1 score by +4.8 % and the professions accuracy by +4.9 % over the sentence-level baseline. This may also highlight the importance of source-side document-level context as the LM based approaches do not have access to this. Still, both back-translation and LM fusion can be combined and this yields further improvements: The best performing approach not relying on document-level parallel data is to use both document-level back-translation and then LM fusion with on-the-fly scales, this method achieves on average +6.2 % F1 score on the pronouns and +6.0 % professions accuracy.

Parallel document-level data. The three baselines trained on parallel document-level data perform much better than the sentence-level baseline: The document-level baselines score on average +18.0 % better on the pronouns F1 score and +4.2 % better on the professions accuracy than their sentence-level counterparts. In addition, the systems trained on parallel documents also perform better than the sentence-level systems with additional monolingual documents in almost all cases. This concludes that on these three tasks, having access to parallel document-level data is much more effective than utilizing monolingual document-level data, even though our monolingual corpora are much larger than the parallel ones.

Method	contrastive pronoun acc.		
	News	Subtitles	e-Comm.
sentence-level baseline	49.0	46.4	46.1
(Jean, 2020; Sugiyama, 2021)[5]	53.4	48.8	47.4
(Jean, 2020)[6]	49.2	46.8	45.5
LM: static	55.2	49.5	48.5
LM: on-the-fly	55.9	53.4	51.3
LM: auto. learned	53.0	50.1	50.5
LM: re-rank pronouns	65.7	73.9	64.8
back-translation	56.5	57.9	47.3
+ LM: static	57.7	61.6	47.5
+ LM: on-the-fly	57.9	61.1	54.3
+ LM: auto. learned	56.7	59.0	54.1
document-level baseline	**67.9**	**84.0**	n.a.

Table 3: Scoring-based, contrastive accuracies on the pronouns test set (Müller et al., 2018) for the three En→De tasks, reported in percent.

Further including monolingual document-level data to the document-level baselines does not generally give additional improvements. In particular, LM pronoun re-ranking decreases performance in this setting as the MT model itself is already better at predicting the correct pronoun than the LM trained on the document-level monolingual data.

Contrastive scores. Previous work on document-level MT commonly evaluates document-level MT systems using contrastive scoring (e.g., Jean and Cho, 2020; Sugiyama and Yoshinaga, 2021). As a direct comparison, we report the contrastive accuracies on the pronouns test set in Table 3. The trend is often similar to the translation-based metrics in Table 2, however scoring-based improvements are much more pronounced. Our experiments also show that strong contrastive accuracies do not necessarily lead to improvements on the generated hypothesis. For example, on the News task, the contrastive scores of the LM pronoun re-ranking approach and the document-level baseline are similar but their translation-based scores differ strongly (c.f. Table 2).

6.4.3 Computational Cost

We have shown that both the on-the-fly scales and the automatically learned scales improve document-targeted scores over static scores obtained via grid search. Another downside of grid search is that the tuning process is quite expensive. In Table 4, we illustrate that a grid search with 11^3 parameters (as is used in this work) on a single GPU can easily take multiple days. The on-the-fly scales do not

Method	Time	
	Preparation	Search
LM: static	7187 min	5.4 min
LM: on-the-fly	0 min	6.5 min
LM: auto. learned	8.3 min	5.4 min

Table 4: Total time necessary to tune different fusion scale variants on a single GPU, as well as the time spent during translation. We measure the time used to translate the News validation set.

LM	perplexity		contrastive acc.	
	news	e-comm.	pron.	proff.
NewsCrawl	17.0	44.5	62.8	**63.4**
LLaMA	**9.2**	**11.8**	**80.0**	62.3

Table 5: Comparing the small in-domain LM trained on NewsCrawl against the LLM LLaMA.

require any preparation time as they are obtained entirely during search, in which the overhead is small. The automatically scales on the other hand can be learned in just a few minutes and do not have any overhead in decoding.

6.4.4 Large Language Model Integration

Recently, large language models (LLMs) which are trained on large corpora and long context sizes received a lot of attention (e.g., Brown et al., 2020; Touvron et al., 2023). In particular, they have also been able to perform document-level MT (Zhang et al., 2023; Hendy et al., 2023; Karpinska and Iyyer, 2023; Wang et al., 2023). This raises the natural question whether LLMs can improve document-level LM fusion.

We experiment on the News task and compare our own small LM with 35M parameters trained on 2.2B tokens from the in-domain German NewsCrawl corpus against the 13B parameter version of LLaMA (Touvron et al., 2023), which was trained on a total of 1000B tokens. LLaMA's training data includes various domains and languages. Only a small fraction of its data is German. The small LM provides two sentences context while we query the LLM with 200 tokens context. We re-train our MT model and the small LM using the LLaMA tokenizer. This leads to slightly worse performance compared to our previous experiments as the LLaMA tokenization was learned on general-domain English data. For decoding we use a beam size of 4.

Table 5 shows the perplexities of both LMs and their contrastive scores on the document-targeted test sets[7]. Both LMs use the same vocabulary and thus their perplexities are comparable. Because it is in general unclear whether test sets are or are not included in LLM training data, we also include the e-Commerce test set which was translated by ourselves for the purpose of cross-validation. On both test sets, the LLM perplexities are much better than the ones of the small in-domain LM. LLaMA's contrastive scores are also much better on the pronouns test set.

LM Fusion		news		e-Commerce	
LM	Scales	BLEU	COMET	BLEU	COMET
(none)	-	31.2	81.3	13.6	70.5
NewsCrawl	static	33.2	83.0	14.4	72.5
	on-the-fly	33.2	82.8	14.3	72.2
LLaMA	static	**34.6**	**84.2**	**16.5**	**75.0**
	on-the-fly	33.4	83.9	13.7	72.9

Table 6: Comparing fusion with a small LM and a LLM on general test sets.

Table 6 shows the performance of LM fusion with the two LMs in BLEU and COMET. Both LMs notably improve translation, but the LLM translation quality is best. Fusion with LLaMA yields +3.4 % absolute improvements on the in-domain test set. Improvements on the e-Commerce test set are similar, indicating that the gains are not an effect of data leakage of the test set into the training data. While the on-the-fly scales and the static scales perform similarly for the small LM, on-the-fly scales do not perform as well for the LLM.

The improvements measured on the in-domain test sets are likely not because of document-level context but rather due to the increased amount of data. Therefore, we continue our evaluation with the document-targeted scores. Table 8 depicts the results. On these metrics, the LLM outperforms the small LM by an even larger margin. In general, the improvements are correlated to their contrastive scores (c.f. Table 5).

6.5 Extended Analysis on Automatically Learned Fusion Scales

In our experiments we use the validation set of the News task to find the best working methods. We share some insights in the following.

How are the automatically learned scales distributed? Figure 1 shows the distribution of the au-

[7] The professions test set was released without target-side context, which we therefore created ourselves by translating the source-side context with a commercial MT system.

Fusion Scales Learning			λ	valid set		doc.-targeted	
Scales	Crit.	Train Set		BLEU	COMET	pron.	proff.
none	-	-	0.0	24.5	80.9	45.1	65.9
subword-agnostic	grid search	valid set	0.40	25.4	81.8	46.5	65.1
	CE	valid set	0.34	25.5	81.7	46.4	65.0
		synthetic	0.46	25.3	81.6	**47.1**	65.2
subword-dependent	CE	valid set	-	**26.6**	**81.8**	46.1	65.0
		synthetic	-	25.4	81.6	46.9	65.4

Table 7: Automatically learning subword-dependent and -agnostic fusion scales on the News task. We employ the restriction $\lambda_0 := 1$, $\lambda := \lambda_1 = \lambda_2$.

LM Fusion		document-targeted		
LM	Scales	pron.	proff.	form.
(none)	-	44.5	65.7	33.4
NewsCrawl	static	46.3	66.3	34.7
	on-the-fly	47.4	66.7	34.3
LLaMA	static	**51.6**	66.9	**36.1**
	on-the-fly	48.2	**68.9**	35.2

Table 8: Fusion with a small LM against a LLM, reporting the translation-based scores on the document-targeted test sets.

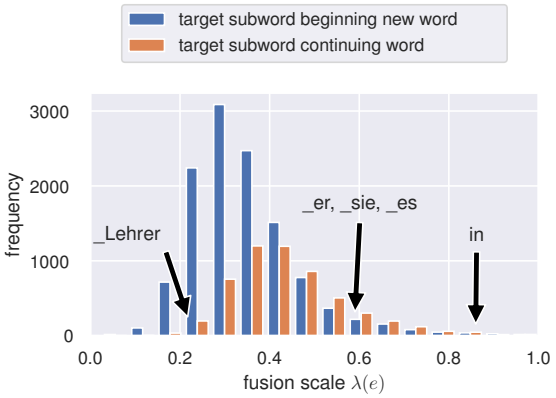


Figure 1: Distribution of the automatically learned LM fusion scales for different target-side subwords on the News task. Subwords for which document-level context is often necessary, such as the German pronouns '_er', '_sie', '_es', and the suffix 'in' marking female professions, have learned higher scales than nouns like '_Lehrer'.

tomatically learned scales for the News task. The learned LM scale of subwords that continue another subword are in general higher than the ones that begin a new word. This is intuitive as continuing a subword is an LM task while beginning a new word requires information about the source sentence.

How much data is needed for automatically learning scales? The static fusion scales are usually tuned on a small validation set via grid search. Table 7 shows that it is also possible to use automatic differentiation to learn static scales only on the validation set. The automatically learned subword-agnostic scales have similar values as the ones tuned via grid search and therefore also their translation performance is similar. Learning subword-dependent scales automatically on the validation set on the other hand improves performance on this set, but does not generalize which indicates overfitting.

7 Conclusions

This work presents multiple extensions to document-level LM fusion, a technique of utilizing document-level monolingual data for context-aware MT. In comparison to existing work, our extensions significantly improve discourse-modeling across four MT tasks and furthermore are computationally more efficient. We conduct evaluations against two baselines: document-level back-translation and a task-specific LM re-ranking method. Despite our extensions, back-translation in general still outperforms document-level LM fusion. Nevertheless back-translation can be effectively combined with LM fusion, further improving translation performance. On very specific test sets, the LM re-ranking performs best. However, our experiments also show that systems trained on document-level parallel data outperform the best systems trained with monolingual documents only.

Finally, this work is the first to explore document-level LM fusion with LLMs. First findings demonstrate that fusion with an LLM outperforms a small LM trained on in-domain data and open the path for future investigations.

Limitations

The experiments in this work were limited to four MT tasks, from which two are low-resource and three are translating from English into German. Apart from the experiments with the LLM, we did not conduct any experiments on a large-scale dataset of multi-domain monolingual documents. The LLM in our experiments only has 7B parameters, while much larger LLMs exist (e.g., Touvron et al., 2023).

Further, our work focuses only on one specific architecture for document-level MT and uses only two sentences target-side context. Various other architectures exist and may entail different properties. This work further does not investigate the behavior of larger translation models.

Another limitation lies in the evaluation of document-level MT models. The document-level targeted metrics we used are all reference-based and limited to the translation of pronouns, gender-referring professions or salutation forms. Other discourse phenomena like e.g. cohesion exist (Maruf et al., 2022) but were not studied in our work. It is unclear how well automated metrics actually correlate with the actual document-level translation quality (Currey et al., 2022), and this work did not perform any qualitative analysis.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz-Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. Paracrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4555–4567. Association for Computational Linguistics.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1304–1313. Association for Computational Linguistics.

Alexandre Berard, Ioan Calapodescu, and Claude Roux. 2019. Naver labs europe's systems for the WMT19 machine translation robustness task. In *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1*, pages 526–532. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Conference on Spoken Language Translation, IWSLT 2017, Tokyo, Japan, December 14-15, 2017*, pages 2–14. International Workshop on Spoken Language Translation.

Linqing Chen, Junhui Li, Zhengxian Gong, Boxing Chen, Weihua Luo, Min Zhang, and Guodong Zhou. 2021a. Breaking the corpus bottleneck for context-aware neural machine translation with cross-task pre-training. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2851–2861. Association for Computational Linguistics.

Linqing Chen, Junhui Li, Zhengxian Gong, Xiangyu Duan, Boxing Chen, Weihua Luo, Min Zhang, and Guodong Zhou. 2021b. Improving context-aware neural machine translation with source-side monolingual documents. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 3794–3800. ijcai.org.

Anna Currey, Maria Nadejde, Raghavendra Reddy Pappagari, Mia Mayer, Stanislas Lauly, Xing Niu, Benjamin Hsu, and Georgiana Dinu. 2022. Mt-geneval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In *Proceedings of the 2022 Conference on Empirical Methods*

in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022, pages 4287–4299. Association for Computational Linguistics.

Miquel Esplà-Gomis, Mikel L. Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. Paracrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks, MT-Summit 2019, Dublin, Ireland, August 19-23, 2019*, pages 118–119. European Association for Machine Translation.

Thierry Etchegoyhen and Harritxu Gete. 2020. To case or not to case: Evaluating casing methods for neural machine translation. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 3752–3760. European Language Resources Association.

Eva Martínez Garcia, Carles Creus, and Cristina España-Bonet. 2019. Context-aware neural machine translation decoding. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation, DiscoMT@EMNLP 2019, Hong Kong, China, November 3, 2019*, pages 13–23. Association for Computational Linguistics.

Yusser Al Ghussin, Jingyi Zhang, and Josef van Genabith. 2023. Exploring paracrawl for document-level neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 1296–1302. Association for Computational Linguistics.

Çaglar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *CoRR*, abs/1503.03535.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? A comprehensive evaluation. *CoRR*, abs/2302.09210.

Christian Herold, Yingbo Gao, Mohammad Zeineldeen, and Hermann Ney. 2023. Improving language model integration for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*. Association for Computational Linguistics.

Christian Herold and Hermann Ney. 2023a. Improving long context document-level machine translation. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse: CODI 2023, Toronto, Canada, July 13-14, 2023*. Association for Computational Linguistics.

Christian Herold and Hermann Ney. 2023b. On search strategies for document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*. Association for Computational Linguistics.

Jingjing Huo, Christian Herold, Yingbo Gao, Leonard Dahlmann, Shahram Khadivi, and Hermann Ney. 2020. Diving deep into context-aware neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation, WMT@EMNLP 2020, Online, November 19-20, 2020*, pages 604–616. Association for Computational Linguistics.

Sébastien Jean, Ankur Bapna, and Orhan Firat. 2019. Fill in the blanks: Imputing missing sentences for larger-context neural machine translation. *CoRR*, abs/1910.14075.

Sébastien Jean and Kyunghyun Cho. 2020. Log-linear reformulation of the noisy channel model for document-level neural machine translation. In *Proceedings of the Fourth Workshop on Structured Prediction for NLP@EMNLP 2020, Online, November 20, 2020*, pages 95–101. Association for Computational Linguistics.

Sébastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine translation benefit from larger context? *CoRR*, abs/1704.05135.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. Blonde: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1550–1565. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1*, pages 225–233. Association for Computational Linguistics.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. *CoRR*, abs/2304.03245.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers, MTSummit 2005, Phuket, Thailand, September 13-15, 2005*, pages 79–86.

Shaohui Kuang and Deyi Xiong. 2018. Fusing recency into neural machine translation with an inter-sentence gate model. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 607–617. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 66–75. Association for Computational Linguistics.

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human-machine parity in language translation. *J. Artif. Intell. Res.*, 67:653–672.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? A case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4791–4796. Association for Computational Linguistics.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8:726–742.

António V. Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, EAMT 2020, Lisboa, Portugal, November 3-5, 2020*, pages 225–234. European Association for Machine Translation.

Sameen Maruf and Gholamreza Haffari. 2018. Document context neural machine translation with memory networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1275–1284. Association for Computational Linguistics.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2022. A survey on document-level neural machine translation: Methods and evaluation. *ACM Comput. Surv.*, 54(2):45:1–45:36.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2947–2954. Association for Computational Linguistics.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 61–72. Association for Computational Linguistics.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. 2019. When does label smoothing help? In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 4696–4705.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *CoRR*, abs/1904.01038.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 186–191. Association for Computational Linguistics.

Matt Post and Marcin Junczys-Dowmunt. 2023. Escaping the sentence-level paradigm in machine translation. *CoRR*, abs/2304.12959.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2685–2702. Association for Computational Linguistics.

Fahimeh Saleh, Alexandre Berard, Ioan Calapodescu, and Laurent Besacier. 2019. Naver labs europe's systems for the document-level generation and translation task at WNGT 2019. In *Proceedings of the 3rd Workshop on Neural Generation and Translation@EMNLP-IJCNLP 2019, Hong Kong, November 4, 2019*, pages 273–279. Association for Computational Linguistics.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. Ccmatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6490–6500. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Felix Stahlberg, Danielle Saunders, Adrià de Gispert, and Bill Byrne. 2019. Cued@wmt19: Ewc&lms. In *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1*, pages 364–373. Association for Computational Linguistics.

Amane Sugiyama and Naoki Yoshinaga. 2019. Data augmentation using back-translation for context-aware neural machine translation. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation, DiscoMT@EMNLP 2019, Hong Kong, China, November 3, 2019*, pages 35–44. Association for Computational Linguistics.

Amane Sugiyama and Naoki Yoshinaga. 2020. Context-aware decoder for neural machine translation using a target-side document-level language model. *CoRR*, abs/2010.12827.

Amane Sugiyama and Naoki Yoshinaga. 2021. Context-aware decoder for neural machine translation using a target-side document-level language model. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 5781–5791. Association for Computational Linguistics.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. Rethinking document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3537–3548. Association for Computational Linguistics.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation, DiscoMT@EMNLP 2017, Copenhagen, Denmark, September 8, 2017*, pages 82–92. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. Context-aware monolingual repair for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 877–886. Association for Computational Linguistics.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1264–1274. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. *CoRR*, abs/2304.02210.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Lei Yu, Laurent Sartran, Wojciech Stokowiec, Wang Ling, Lingpeng Kong, Phil Blunsom, and Chris Dyer. 2020. Better document-level machine translation with bayes' rule. *Trans. Assoc. Comput. Linguistics*, 8:346–360.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. *CoRR*, abs/2301.07069.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 533–542. Association for Computational Linguistics.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating BERT into neural machine translation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Appendix

A.1 Model Training

Data. The *News En→De* task comprises 330k parallel sentences from NewsCommentary v14[8], which we combine with document-level monolingual data from NewsCrawl[9] (70M sentences[10]). Our *Subtitles En→De* data consists of a total of 39M monolingual movie show subtitles from OpenSubtitles, from which a subset of 22.5M sentences has been aligned to English sentences and forms our parallel training data (Lison et al., 2018). For *TED En→It* we use 230k parallel sentences from scientific TED talks released as part of the IWSLT17 multilingual task (Cettolo et al., 2017) which we combine with 2.2M sentences of talks from the European parliament (Koehn, 2005). Finally, the *e-Commerce En→De* task is about translating item descriptions from e-Commerce listings. We use 326M parallel sentences of out-of-domain parallel training data from the ParaCrawl v9 corpus (Esplà-Gomis et al., 2019) which we combine with 128k parallel sentences in-domain data. The monolingual data was sampled from item descriptions and is entirely in-domain (119M sentences).

The sizes of our training corpora are shown in Table 10.

On each task, we use a validation set for selecting the best checkpoint, tuning the fusion scales and for finding which method works best. For the final comparison in Table 1 we then report on an unseen test set of the same domain.

The News validation set is `newstest2015`, and `newstest2018` as test set. For Subtitles, our validation and test sets were sampled from the training corpus. The precise document IDs for the validation set are: 1995/254, 1997/165, 2000/313, 2002/461, 2005/441, 2007/781, 2010/273, 2012/757, 2015/1488, 2017/525 for the validation set; and for our test set: 1997/310, 2002/40, 2007/189, 2012/1085, 2017/644. The test set is the same as used in Huo et al. (2020). For TED, we concatenate `dev2010` and `tst2010` and use `tst2017.mltlng` as test set. For e-Commerce, we create the validation and test set ourselves by translating English e-Commerce item descriptions into German: Our validation set comprises 85 documents (2882 sentences) and the test set 100

[8] https://data.statmt.org/news-commentary/v14/

[9] https://data.statmt.org/news-crawl/

[10] To reduce training time, our back-translation experiments on this task utilize only the first 2M sentences.

Data	News			Subtitles			TED		e-Commerce		
	test	pron.	proff.	test	pron.	proff.	test	proff.	test	pron.	proff.
parallel data	129.7	125.7	168.6	**26.6**	**36.0**	**103.2**	**47.6**	**114.8**	61.7	**44.2**	**52.6**
monolingual data	**97.1**	**94.9**	**161.3**	27.8	36.6	117.4	75.4	116.1	**50.8**	48.7	57.5

Table 9: Perplexities of sentence-level LMs trained on equal amount of target-side data.

Task	Data	docs	sents	words
News	parallel	8.5k	330k	7.4M
	mono.	3M	70M	1.0B
Subtitles	parallel	30k	22.5M	136M
	mono.	47k	39M	223M
TED	parallel	1.9k	230k	3.7M
	mono.	6k	2.2M	54.6M
e-Commerce	parallel	n.a.	326M	9.6B
	mono.	1.5M	119M	3.1B

Table 10: Training data statistics.

documents (2520 sentences).

As the pronouns test set (Müller et al., 2018) was extracted from the OpenSubtitles corpus, we remove these sentences from the Subtitles training data. The professions test set (Currey et al., 2022) was curated from Wikipedia articles and is not part of our training corpora.

Models. We train the News, Subtitles and TED models with a shared embedding and projection matrix. Th resulting MT models for News and Subtitles have 60M parameters, 51M parameters for TED and 90M for e-Commerce. For model training we use eight Tesla V100-SXM2-32GB GPUs. Training the baselines takes approximately 7h for News, 21h for Subtitles, 5h for TED, and 30h for e-commerce. Due to resource constraints, we report only a single run for each experiment.

Optimization. For optimization we use Adam (Kingma and Ba, 2015) and a batch size of 22k subwords. The low-resource MT models (News, TED) are trained for 100k update steps with 30 % dropout, 20 % label smoothing and weight decay, while the high-resource models (Subtitles, e-Commerce) are trained for 300k updates with 10 % dropout, 10 % label smoothing and no weight decay.

A.2 Domain Effects

In an effort to estimate how well the domain of the training data matches the test sets, we train LMs on the target-side part of the parallel and the monolingual training data. Within each task, the LMs are trained with the same parameters and the same vocabulary. We then report the perplexities on the task-specific test sets and the document-targeted challenge sets in Table 9.

For News, the monolingual data is more in-domain for all test sets. Similarly the domain of the e-Commerce monolingual data is closer to the task-specific test set. For Subtitles, the domains of parallel and monolingual data are more or less equal and on TED, the monolingual data is slightly out-of-domain.

This domain effect explains the improvements in BLEU and COMET on the task-specific test sets that we reported in Table 1 on News and on e-Commerce.

Approach	valid set		doc.-targeted	
	BLEU	COMET	pron.	proff.
baseline	24.5	80.9	45.1	**65.9**
LM fusion	24.8	81.2	45.0	65.8
+ (Jean, 2020; Sugiyama, 2021)[5]	24.9	81.2	46.0	65.0
+ ILM: separate	**25.8**	**82.1**	**47.3**	65.2
+ ILM: $h = 0$	25.5	82.0	43.8	64.7
+ ILM: mini self-att.	**25.8**	**82.1**	44.6	65.1

Table 11: Document-level LM fusion (a) without subtracting any LM, (b) subtracting the sentence-level probabilities of the external LM (Jean and Cho, 2020; Sugiyama and Yoshinaga, 2021), and (c) subtracting different approximations of the internal LM (ILM) learned by the MT model, reported on the News task.

A.3 Comparing Internal Language Model Estimations

Herold et al. (2023) propose several ways of approximating the internal LM learned implicitly by the MT model in the context of sentence-level MT. We evaluate three of their approaches for document-level LM fusion and compare them against the existing document-level LM fusion approach that subtracts the sentence-level probabilities of the external LM (Jean and Cho, 2020; Sugiyama and Yoshinaga, 2021). Table 11 shows the results: Subtracting the internal LM substantially improves LM fusion over existing work. Estimating it by training a separate LM on the same data as the MT model works best.

Fusion Scales		valid set		doc.-targeted	
Approach	Restriction	BLEU	COMET	pron.	proff.
none	-	24.5	80.9	45.1	65.9
static	-	**25.8**	**82.1**	47.3	65.2
	$\lambda_0 = 1, \lambda_1 = \lambda_2$	25.4	81.8	46.5	65.1
on-the-fly	-	22.3	78.3	43.4	69.3
	$\lambda_0 = 1, \lambda_1 = \lambda_2$	25.6	81.8	**48.0**	65.5
auto. learned	-	24.8	80.7	44.7	**69.4**
	$\lambda_0 = 1, \lambda_1 = \lambda_2$	25.3	81.5	46.7	64.9

Table 12: LM fusion with an imposed restriction on the search space of the fusion scales $\lambda_0, \lambda_1, \lambda_2$, reported on the News task.

Data		valid set		doc.-targeted	
parallel	mono.	BLEU	COMET	pron.	proff.
sent.	-	24.5	80.9	45.1	65.9
sent.	sent.	27.0	**83.2**	46.7	65.7
sent.	doc.	26.9	82.4	47.8	**80.7**
pseudo-doc.	doc.	**27.1**	83.0	**48.7**	80.5

Table 13: Effect of back-translation on the News task.

A.4 Fusion Scale Restrictions

The three LM fusion scales $\lambda_0, \lambda_1, \lambda_2$ in Equation 1 balance the contribution of the MT model and the two LMs. In our experiments the optimal scales usually lie at $\lambda_0 \approx 1$ and $\lambda_1 \approx \lambda_2$. This is plausible as the internal LM (λ_2) should neutralize the external LM (λ_1) to the same degree. For the non-static fusion scales however, we find that searching over the three-dimensional search space of independent $\lambda_0, \lambda_1, \lambda_2$ finds unintuitive scale combinations and that this causes bad performance. Therefore in our experiments we restrict the search space of fusion scales to the one-dimensional slice where $\lambda_0 = 1$ and $\lambda_1 = \lambda_2$. Table 12 gives a direct comparison.

A.5 Document-level Back-translation

In Table 13 we compare back-translation using document-level data against sentence-level back-translation.

Sentence- and document-level back-translation gives the same performance improvements in BLEU and COMET, however only back-translation on document-level improves the document-targeted metrics. For document-level back-translation we find that creating pseudo-documents from the parallel data is necessary to achieve the same BLEU and COMET scores as sentence-level back-translation.

ChatGPT MT: Competitive for High- (but not Low-) Resource Languages

Nathaniel R. Robinson[1,2*] **Perez Ogayo**[1*] **David R. Mortensen**[1] **Graham Neubig**[1]
[1]Language Technologies Institute, Carnegie Mellon University
[2]Center for Language and Speech Processing, Johns Hopkins University
nrobin38@jhu.edu, {aogayo, dmortens, gneubig}@cs.cmu.edu
* Authors contributed equally

Abstract

Large language models (LLMs) implicitly learn to perform a range of language tasks, including machine translation (MT). Previous studies explore aspects of LLMs' MT capabilities. However, there exist a wide variety of languages for which recent LLM MT performance has never before been evaluated. Without published experimental evidence on the matter, it is difficult for speakers of the world's diverse languages to know how and whether they can use LLMs for their languages. We present the first experimental evidence for an expansive set of 204 languages, along with MT cost analysis, using the FLORES-200 benchmark. Trends reveal that GPT models approach or exceed traditional MT model performance for some high-resource languages (HRLs) but consistently lag for low-resource languages (LRLs), under-performing traditional MT for 84.1% of languages we covered. Our analysis reveals that a language's resource level is the most important feature in determining ChatGPT's relative ability to translate it, and suggests that ChatGPT is especially disadvantaged for LRLs and African languages.

1 Introduction

Despite the majority of the world's languages being low-resource, current MT systems still perform poorly on them or do not include them at all. Some commercial systems like Google Translate[1] support a number of LRLs, but many systems do not support any, and in either case the majority of LRLs are largely neglected in language technologies.

In recent years, generative LLMs have shown increasingly impressive translation abilities (Radford et al., 2019; Brown et al., 2020). Even more recently, LLM tools like ChatGPT have become popular and accessible to end users. This marks an important shift, since a majority of LLM users are now consumers rather than researchers. The prospect of LLM translation is exciting, since theoretically, generative LLMs could support more languages than commercial systems like Google's.[2] But only beginning steps have been made to test this hypothesis. While some studies outlined in §4 have evaluated MT with recent LLMs, evaluation is still lacking for many languages. This brings up important questions, such as: *Can end users in need of MT for a variety of languages use ChatGPT? Are ChatGPT and other LLMs reliable translators? For which languages are they reliable?* Initially we hypothesize that LLMs translate HRLs better than LRLs. But due to limited information about the training data and methods for powerful LLMs like ChatGPT (**GPT-3.5** and variants) and GPT-4, hypotheses like this must be experimentally verified.

We significantly expand experimental verification for such hypotheses by testing ChatGPT's performance on the FLORES-200 benchmark (NLLB Team et al., 2022), containing 204 language varieties. We emphasize that, rather than optimizing LLM MT for a few languages, we focus on helping end users of various language communities know how and when to use LLM MT. We expect that our contributions may benefit both direct end users, such as LRL speakers in need of translation, and indirect users, such as researchers of LRL translation considering ChatGPT to enhance specialized MT systems. In summary, we contribute:

1. MT scores on 203 languages for ChatGPT and comparisons with GPT-4, Google Translate, and NLLB (NLLB Team et al., 2022)
2. Evidence that LLMs are competitive with traditional MT models for many HRLs but lag for LRLs (with baselines outperforming ChatGPT on 84.1% of languages evaluated)
3. Evidence that few-shot prompts offer

[1]https://translate.google.com

[2]Google Translate currently supports only 133 languages with systems deemed high enough quality for deployment.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 392–418
December 6–7, 2023.

marginal benefits for LLM translation

4. A decision tree analysis of language features' correlation with LLM effectiveness in MT, suggesting ChatGPT is especially disadvantaged for LRLs and African languages

5. A cost comparison across MT systems

Our experiments are motivated by the interests of LLM users speaking a variety of languages. In addition to evaluating a large language set (§3), we chose to analyse language features (§3.4), to draw generalizations for even more LRL speakers. We compare MT costs because they impact end users (§3.7). We keep ChatGPT central to our analyses because of its current popularity among consumers.

2 Methodology

We used data for 204 language varieties from FLORES-200 (NLLB Team et al., 2022). We used the 1012 *devtest* sentences for our main experiments and the 997 *dev* sentences for follow-up experiments. We queried the OpenAI API[3] to translate our test set from English into the target languages. We explored ENG→X translation only because the FLORES-200 English data was taken from Wikipedia. Thus OpenAI's GPT models were likely trained on those exact English sentences, making fair X→ENG evaluation infeasible.

2.1 Experimental setup

We evaluated ChatGPT's (`gpt-3.5-turbo`) MT for our full language set. We compared with NLLB-MOE (NLLB Team et al., 2022) as our baseline, as it is the current state-of-the-art open-source MT model that covers such a wide variety of languages. NLLB is a discriminative transformer trained on supervised bi-text data (the traditional MT paradigm). We obtained scores for NLLB outputs of ENG→X translation into 201 of the language varieties in our set (as reported by NLLB Team et al. (2022)).

We used both zero- and five-shot prompts for ChatGPT MT. (See §2.3.) Previous studies (Hendy et al., 2023; Gao et al., 2023; Moslem et al., 2023; Brown et al., 2020; Zhu et al., 2023) suggest that few-shot prompts produce slightly (albeit not consistently) better translations. But zero-shot prompts are more convenient and affordable for users.

We also compare with results for subsets of our selected languages from two other MT engines. Google Translate API was an important baseline for our analysis because it is popular among end users. We also included it to represent commercial MT systems in our study. Because Google's API does not support all 204 of the FLORES-200 languages, we obtained results only for the 115 non-English languages it supports.

Lastly, we obtained MT results from GPT-4, since it is a popular LLM and has been shown to outperform ChatGPT on MT (Jiao et al., 2023; Wang et al., 2023). Because the cost of GPT-4 use exceeds that of ChatGPT by 1900%, our resources did not permit its evaluation on all 203 non-English languages. Instead we selected a 20-language subset by picking approximately every 10th language, with languages sorted by chrF++ differentials between ChatGPT and NLLB ($chrf_{GPT} - chrf_{NLLB}$). We chose this criterion in order to have 20 languages with a range of relative ChatGPT performance and a variety of resource levels. We used only five-shot prompts for GPT-4.

2.2 Implementation details

We conducted all LLM experiments with `gpt-3.5-turbo` (ChatGPT) and `gpt-4-0613` (GPT-4). We used `top_p` 1, `temperature` 0.3, `context_length` −1, and `max_tokens`[4] 500.

To evaluate the outputs, we used:[5]

spBLEU: BLEU (Papineni et al., 2002) is standard in MT evaluation. We find spBLEU scores (Goyal et al., 2022) via sacreBLEU (Post, 2018) with the SPM-200 tokenizer (NLLB Team et al., 2022).

chrF2++: We use sacreBLEU's implementation of chrF++ (Popović, 2017). We adopt it as our main metric, as it overcomes some of BLEU's weaknesses, and refer to it as *chrF* for brevity.

2.3 Zero- and few-shot prompts

Previous works (Gao et al., 2023; Jiao et al., 2023) investigated LLM prompting to optimize MT performance. We adopt Gao et al. (2023)'s recommended prompts for both zero- and few-shot MT (Table 1). We are interested in multiple n-shot prompt settings because, as mentioned in §2.1, they

[3] `https://platform.openai.com`

[4] Although some languages had higher token counts than others (see §3.4), we found that adjusting `max_tokens` had a minimal effect on MT performance. We thus decided to maintain the same value of `max_tokens` across all languages for experimental consistency.

[5] We excluded learned MT metrics like COMET (Rei et al., 2020) and BLEURT (Sellam et al., 2020), since they do not support many LRLs.

Shot	Prompt
zero	This is an English to [TGT] translation, please provide the [TGT] translation for this sentence. Do not provide any explanations or text apart from the translation. [SRC]: [src-sentence] [TGT]:
five	This is an English to [TGT] translation, please provide the [TGT] translation for these sentences: [SRC]: [src-sentence] [TGT]: [tgt-sentence] [SRC]: [src-sentence] [TGT]: [tgt-sentence] [SRC]: [src-sentence] [TGT]: [tgt-sentence] [SRC]: [src-sentence] [TGT]: [tgt-sentence] [SRC]: [src-sentence] [TGT]: [tgt-sentence] Please provide the translation for the following sentence. Do not provide any explanations or text apart from the translation. [SRC]: [src-sentence] [TGT]:

Table 1: Prompts used for zero- and five-shot settings

present different benefits to LLM users. We explored zero-shot (no in-context example), one-shot (1 example), and five-shot (5 examples). We employed both zero- and five-shot prompts in our main experiments over 203 languages, and we analyzed all three n-shot settings for a subset of languages on FLORES-200 *dev* sets.

The languages in FLORES-200 represent 22 language families. To experiment with multiple n-shot settings, we selected one language from each of the 12 families containing at least two members in the set. We chose four HRLs ($\geq$1M Wikipedia pages[6]), four LRLs (25K-1M pages), and four extremely LRLs ($\leq$25K pages). These languages also employ a variety of scripts. See Table 2.

Language	Code	Family	Script	Wiki. #
French	fra	Indo-European	Latn	12.7M
Chinese	zho	Sino-Tibetan	Hans	7.48M
Turkish	tur	Turkic	Latn	2.48M
Finnish	fin	Uralic	Latn	1.46M
Tamil	tam	Dravidian	Taml	496K
Tagalog	tgl	Austronesian	Latn	239K
Kiswahili	swh	Niger-Congo	Latn	167K
Amharic	amh	Afroasiatic	Ethi	46.2K
Santali	sat	Austroasiatic	Olck	20.0K
Lao	lao	Kra-Dai	Laoo	14.0K
Papiamento	pap	Creole	Latn	6.84K
Luo	luo	Nilo-Saharan	Latn	0

Table 2: Diverse subset of languages experiments with few-shot settings. **Wiki. #** is the number of Wikipedia pages in the language.

[6]Throughout the paper we use the "Total pages" count from https://en.wikipedia.org/wiki/List_of_Wikipedias, accessed 7 August 2023, as a proxy for the resource level of a language.

	#langs.	avg. chrF	avg. BLEU
ChatGPT (0-shot)	203	32.3	16.7
ChatGPT (5-shot)	203	33.1	17.3
GPT-4	20	44.6	24.6
NLLB	201	45.3	27.1
Google	115	**52.2**	**34.6**

Table 3: Languages evaluated, average chrF, and average BLEU for each MT system. Best scores are **bold**.

3 Results and Analysis

3.1 Traditional MT generally beats LLMs

Table 3 shows the number of languages we evaluated for each MT system, as noted in §2.1, with average chrF and BLEU scores across those languages. The best performing model on average was (1) Google, then (2) NLLB, (3) GPT-4, and (4) ChatGPT. Unabridged results are in Table 11 in Appendix A. Supplementary materials can also be browsed on our repository.[7] (Also see the interactive score visualizer on our Zeno browser.[8])

Table 4 shows chrF for the 20 languages evaluated on both LLM systems. Of the 11 languages evaluated on all four systems, Google performed best for 9 of them. Notably, GPT-4 surpassed NLLB in five languages and Google in one[9] (Mesopotamian Arabic, acm_Arab).

On the 20 languages for which we tested it, GPT-4 improved over ChatGPT by 6.5 chrF on average. The standard deviation of performance difference with NLLB ($chrF_{GPT} - chrF_{NLLB}$) was 8.6 for GPT-4, compared with ChatGPT's 12.7 for the same languages, suggesting a more consistent advantage across language directions. GPT-4 offered larger improvements for LRLs, whereas HRL performance plateaued between the LLMs. Previous studies have found GPT-4 improving multilingual capabilities over ChatGPT on a range of tasks (Xu et al., 2023; Zhang et al., 2023; OpenAI, 2023). This may account for its superior MT performance.

Google Translate outperformed all other systems in chrF on 100 of the 115 languages for which we evaluated it, with an average improvement of 2.0 chrF points over the next best system for each language. (See Appendix A for unabridged results.)

[7]https://github.com/cmu-llab/gpt_mt_benchmark

[8]https://hub.zenoml.com/project/cabreraalex/GPT%20MT%20Benchmark

[9]Our language identification analysis in §3.6 and manual inspection suggest that GPT models only output one Arabic variety: Modern Standard Arabic (MSA). It seems the LLMs' high performance on some Arabic varieties is due simply to incidental high token overlap with MSA targets.

Lang.	GPT-4	ChatGPT	Google	NLLB
ssw_Latn	24.1	6.7	-	43.3
sna_Latn	29.2	16.3	**44.4**	43.4
ckb_Arab	33.1	24.8	**47.7**	47.2
mag_Deva	44.6	39.9	-	58.5
ibo_Latn	27.7	16.3	**43.5**	41.4
hau_Latn	40.3	22.4	53.2	**53.5**
pbt_Arab	26.7	21.1	-	39.4
tam_Taml	42.7	34.5	**55.8**	53.7
kat_Geor	41.4	33.5	**51.4**	48.1
gle_Latn	53.0	47.5	**60.1**	58.0
kmr_Latn	34.3	27.4	**40.0**	39.3
war_Latn	54.0	49.5	-	57.4
ajp_Arab	48.4	47.5	-	51.3
lim_Latn	45.1	42.7	-	47.9
ukr_Cyrl	56.3	55.4	**58.6**	56.3
fra_Latn	71.7	71.3	**72.7**	69.7
lvs_Latn	57.3	55.2	-	54.8
ron_Latn	**65.3**	64.2	65.0	61.3
tpi_Latn	**49.5**	39.2	-	41.6
acm_Arab	**46.5**	46.1	-	31.9

Table 4: chrF (↑) scores across models for all languages we used to evaluate GPT-4. Best scores are **bold**. ChatGPT scores here are 5-shot, to compare with GPT-4.

Google's was the best performing MT system overall, though NLLB has broader language coverage.

NLLB outperformed ChatGPT in chrF on 169 (84.1%) of the 201 languages for which we obtained scores for both, with NLLB scoring an average of 11.9 chrF points higher than the better n-shot ChatGPT setting for each language. This trend is corroborated by Zhu et al. (2023). Table 5 has both BLEU and chrF scores from both systems for the five languages with the most negative chrF deltas ($chrF_{GPT} - chrF_{NLLB}$) on top, followed by the five languages with the highest positive deltas on bottom. For many of the subsequent sections of this paper we focus on comparing ChatGPT and NLLB, since we evaluted them on the most languages.

	ChatGPT		NLLB	
Lang.	BLEU	chrF	BLEU	chrF
srp_Cyrl	1.36	3.26	**43.4**	**59.7**
kon_Latn	0.94	8.50	**18.9**	**45.3**
tso_Latn	2.92	15.0	**26.7**	**50.0**
kac_Latn	0.04	2.95	**14.3**	**37.5**
nso_Latn	3.69	16.7	**26.5**	**50.8**
jpn_Jpan	**28.4**	**32.9**	20.1	27.9
nno_Latn	**37.1**	**58.7**	33.4	53.6
zho_Hans	**36.3**	**31.0**	26.6	22.8
zho_Hant	**26.0**	**24.4**	12.4	14.0
acm_Arab	**28.2**	**44.7**	11.8	31.9

Table 5: Lowest (top) and highest (bottom) chrF differences between zero-shot ChatGPT and NLLB. Best scores for each metric in **bold** (with BLEU **blue**).

3.2 ChatGPT under-performs for LRL

Using NLLB Team et al.'s (2022) resource categorization, we find that ChatGPT performs worse on LRLs than HRLs, corroborating findings of previous works (Jiao et al., 2023; Zhu et al., 2023). There is a strong positive correlation between ChatGPT and NLLB chrF scores, but the correlation is higher for HRLs (ρ=0.85) than LRLs (ρ=0.78), indicating that ChatGPT struggles to keep up with NLLB for LRLs.

Figure 1 shows scatter plots where dots represent languages, with ChatGPT's (positive or negative) *relative improvement* over NLLB chrF ($\frac{chrf_{GPT} - chrf_{NLLB}}{chrf_{NLLB}}$) on the y-axis. When languages are grouped by family or script, some trends are apparent (in part because we ordered groups by descending average scores). For example, ChatGPT fairs better with Uralic and Indo-European languages and clearly worse with Niger-Congo and Nilo-Saharan languages. However, the clearest natural correlation appears when languages are grouped by resource level, approximated by number of Wikipedia pages (Figure 1, bottom). Note the *relative improvement* (y-axis) is typically negative since ChatGPT rarely outperformed NLLB.

In the five-shot setting, ChatGPT outperformed NLLB on 47% of the HRLs designated by NLLB Team et al. (2022), but only on 6% of the LRLs. These findings contrast with what is commonly observed in multilingual MT models (Liu et al., 2020; Fan et al., 2020; Siddhant et al., 2022; Bapna et al., 2022; NLLB Team et al., 2022), where LRLs benefit the most. This highlights the need to investigate how decoder-only models may catch up with encoder-decoder models in low-resource applications. It underscores the importance of MT-specialized models when larger multitask models cannot overcome low-resource challenges.

3.3 Few-shot prompts offer marginal improvement

Our main experiments suggested that n-shot setting had only a modest effect on MT performance. We conducted a more concentrated study of n-shot prompts using dev sets for the 12 languages in Table 2. Results in Table 6 show five-shot prompts performing best. For some LRLs, this was simply a result of ChatGPT's failure to model the language. In Santali's case, for example, zero-shot ChatGPT was unable to produce the Ol Chiki script at all. In the five-shot setting, it was able to imitate the script

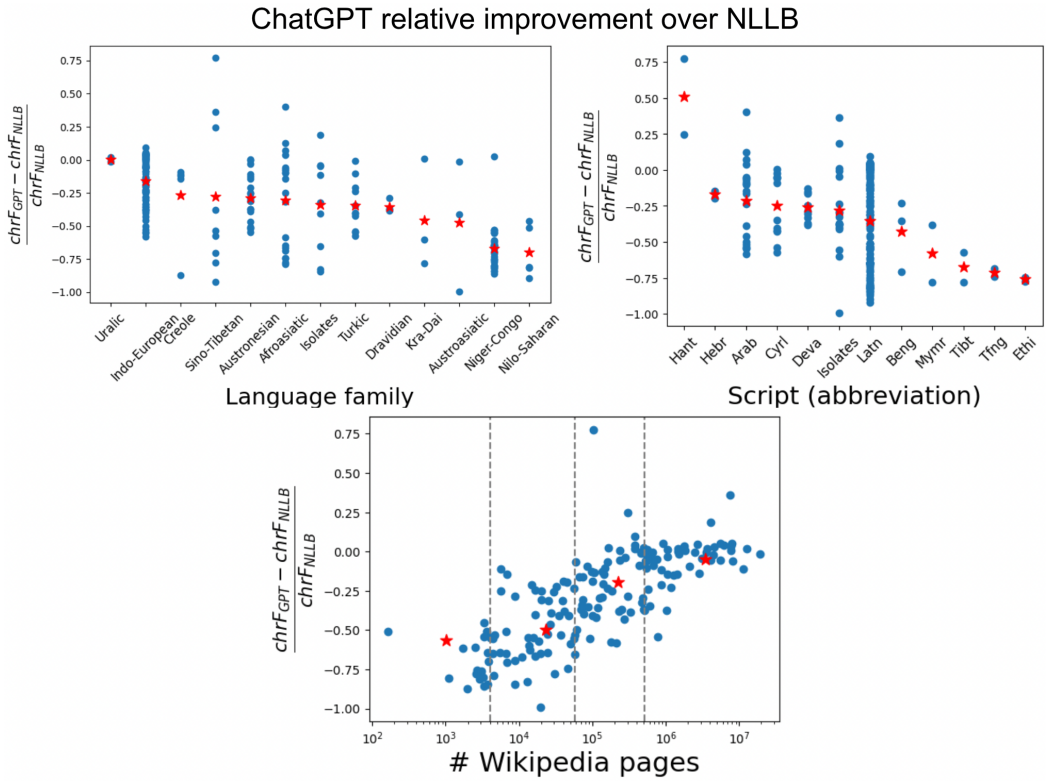


Figure 1: ChatGPT *relative improvement* over NLLB chrF, with languages organized by family, script, and number of Wikipedia pages. Red stars represent averages per group. In the bottom plot, languages are grouped into quartiles of equal size (with dotted lines at the Q1, median, and Q3). More expansive visualizations with language labels for each value can be found in Appendix C.

characters from the context, but without any coherence or accuracy. Excepting Santali as an outlier, five-shot settings offered generally marginal improvements over zero-shot (the most cost-effective of the settings), with an average improvement of only 1.41 chrF across all 12 languages (0.31 if we exclude Santali). Zero-shot prompts actually produced the best chrF score for six of the 12 languages. The one-shot setting performed worst. We noted this trend of few-shot contexts offering only meager and inconsistent improvements throughout our experiments, with five-shot MT improving on zero-shot by only 0.88 average chrF across all 203 language directions. (See Appendix A.)

3.4 Importance of language features

We were interested in which language features determined LLMs' effectiveness compared to traditional MT. Analyzing this may reveal trends helpful to end users deciding which MT system to use, especially if their language is not represented here but shares some of the features we consider. In this section we focus on comparing ChatGPT and NLLB, since we evaluated the most languages with them. We focus on zero-shot ChatGPT, as it is the most common and convenient setting for end users.

	0-shot		1-shot		5-shot	
	BLEU	chrF	BLEU	chrF	BLEU	chrF
fra	55.4	**71.3**	50.4	70.3	55.4	71.2
zho	30.0	29.9	28.2	30.8	**30.7**	**31.1**
fin	34.6	56.6	31.7	56.3	34.6	**56.7**
tur	38.2	**58.6**	34.8	57.6	**38.3**	58.6
tgl	35.9	**60.2**	35.2	59.6	**36.1**	60.1
tam	**13.8**	**35.3**	11.7	34.3	11.9	34.6
swh	39.7	**60.6**	36.0	59.5	**40.0**	60.5
amh	3.4	10.1	3.2	9.6	**3.9**	**10.6**
pap	26.6	51.5	29.3	54.1	**34.8**	**56.1**
lao	4.8	21.6	4.4	20.8	**5.3**	**22.1**
luo	**0.8**	**7.6**	0.2	4.6	0.2	5.2
sat	0.0	0.3	2.2	11.3	**3.0**	**13.8**

Table 6: Three n-shot settings for 12 diverse languages

We encoded each of the 203 languages in our set as a *feature vector*. In these language *feature vectors* we included **four numerical features**: number of Wikipedia pages in the language (`wiki_ct`), size of the language's bi-text corpus in the Oscar MT database[10] (`oscar_ct`) (Abadji et al., 2022), percentage of ASCII characters[11] in the FLORES-

[10] https://oscar-project.org

[11] Percentage of characters with an encoding between 0 and

200 *dev* set for the language (`ascii_percentage`), and average number of tokens per *dev* set sentence in FLORES-200 with ChatGPT's tokenizer (`token_ct`). We also included **two categorical features**: language family (`family`) and script the language was written in (`script`); and **one binary feature**: the FLORES resource designation of the language—with 1 for high-resource and 0 for low-resource (`hi/lo`). Before analysis, we one-hot encoded the two **categorical features** into 48 binary features like `family_Niger-Congo` and `script_Latn`.

We selected `token_ct` as a feature because we observed languages in low-resource scripts having many tokens. For example, ChatGPT's tokenizer encodes multiple tokens for every character in Ol Chiki script. This tendency for GPT models with low-resource scripts has been noted in previous studies (Ahia et al., 2023).

We fit a decision tree with these *feature vectors* to regress on ChatGPT's *relative improvement* over NLLB in chrF ($\frac{chrf_{GPT} - chrf_{NLLB}}{chrf_{NLLB}}$), for each of the 201 languages with NLLB scores. When we used `max_depth` 3, the tree in Figure 2 was learned. Languages are delimited first by `wiki_ct`; then LRLs are separated into Niger-Congo languages and others, while HRLs are delimited by `token_ct`. The only group where ChatGPT beat NLLB is of languages with more than 58,344 Wikipedia pages, fewer than 86 tokens per average sentence, and less than 15.5% ASCII characters. This group contains some East Asian HRLs. The group where ChatGPT was least advantaged contains Niger-Congo languages with fewer than 3,707 Wikipedia pages.

We also fit a random forest regressor with the same features and labels to find feature importance values. Only ten features had importance ≥ 0.01, shown in Table 7. The most important feature by far was `wiki_ct`. (This feature correlates strongly with ChatGPT's *relative improvement*, $\rho = 0.68$.) `family_Niger-Congo` was much more important than any other family feature. No script feature had an importance exceeding 0.01. In general, features for resource level and tokenization were more important than family or script.

ChatGPT has a blind spot not only for Niger-Congo languages, but for African languages in general. Figure 1 shows ChatGPT is least advantaged for the two exclusively African families, Niger-Congo and Nilo-Saharan; and the two exclusively African scripts, Tifinagh (`Tfng`) and Ge'ez (`Ethi`).

feature	importance
`wiki_ct`	0.514
`token_ct`	0.157
`ascii_percentage`	0.104
`family_Niger-Congo`	0.054
`oscar_ct`	0.040
`family_Afroasiatic`	0.025
`family_Indo-European`	0.025
`family_Sino-Tibetan`	0.022
`family_Creole`	0.012
`family_Nilo-Saharan`	0.011

Table 7: Ten most important language features to predict ChatGPT's effectiveness relative to NLLB

3.5 Impact of script

Prior research suggests that ChatGPT output quality is sensitive to language script (Bang et al., 2023). Our own analysis in §3.4 actually suggests that script is the least important language feature in predicting ChatGPT's MT effectiveness. However, differences in performance are clear when comparing scripts used for the same language. Table 8 shows one script typically outperforming the other, by an average of 14.3 chrF points for zero-shot. Five-shot contexts narrowed the gap slightly to 12.0. Although transliteration is a deterministic process for many languages, these performance gaps suggest that ChatGPT has not implicitly learned it as part of a translation task. We hypothesize that ChatGPT's observed sensitivity to script in earlier studies may be particular to the languages and tasks evaluated.

	BLEU		chrF	
Lang.	0-shot	5-shot	0-shot	5-shot
ace_Arab	1.27	2.26	8.41	9.75
ace_Latn	**4.98**	4.35	**19.82**	17.96
arb_Arab	37.60	**37.85**	53.79	**53.81**
arb_Latn	5.33	8.38	22.79	26.92
bjn_Arab	1.96	3.05	10.43	13.24
bjn_Latn	10.96	**12.29**	35.92	**37.98**
kas_Arab	**3.99**	3.30	**15.51**	14.33
kas_Deva	2.31	2.68	12.91	13.91
knc_Arab	0.51	1.06	5.26	4.67
knc_Latn	**2.61**	0.91	**13.38**	8.11
min_Arab	1.56	3.49	10.06	14.88
min_Latn	11.51	**13.07**	36.99	**38.43**
taq_Latn	0.82	0.28	8.18	6.24
taq_Tfng	0.62	**1.37**	5.23	**8.31**
zho_Hans	36.33	**36.51**	31.03	**31.89**
zho_Hant	29.30	30.38	24.82	26.02

Table 8: ChatGPT performance on languages with multiple scripts. Each better scoring script is **bold**.

128, inclusive, using the Python built-in `ord` function

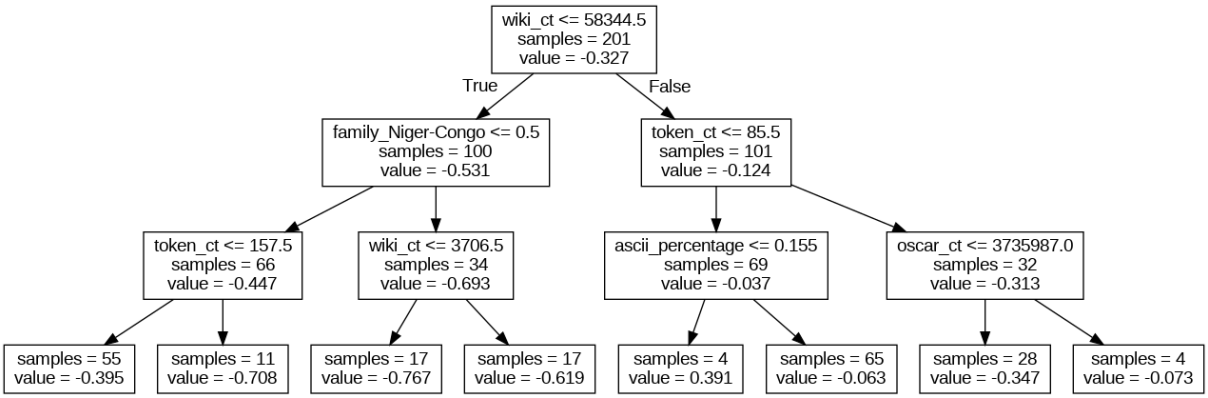


Figure 2: Decision tree predicting ChatGPT *relative improvement* over NLLB chrF, from language features.

3.6 LLMs often get the language wrong

LLMs' performing worse than NLLB may be due in large part to their translating into the wrong language. Using FLORES-200's *dev* data, we trained a logistic regression language identifier for 100 epochs. Language identification accuracies for four of the models we evaluated are in Table 9. Zero-shot ChatGPT only translated on target 72% of the time. This expectedly improved with five-shot prompts, and GPT-4 performed even better, still just shy of NLLB. LLMs' tendency to translate off target is corroborated by Zhu et al. (2023).

model	lang. ID acc.
ChatGPT (0-shot)	72%
ChatGPT (5-shot)	83%
GPT-4 (5-shot)	90%
NLLB	91%

Table 9: Proportion of the time each model translated into the correct target language

3.7 Cost comparison

Our results suggest that GPT-4 is a better translator than ChatGPT. However in considering the needs of MT end users, it would be remiss not to consider the respective costs of the systems evaluated. GPT-4's high cost (roughly 2000% that of ChatGPT's) prohibited us from evaluating it on all FLORES-200 languages. In general, using few-shot prompts for LLMs is more costly than zero-shot prompts, since users are charged for both input and output tokens. And for this same reason, some languages are more costly than others in LLM MT. Previous work has found that Google Translate has associated costs comparable to those of five-shot ChatGPT (Neubig and He, 2023). NLLB is the least expensive system we evaluated.

We estimated cost values for each MT system and language: the expense, in USD, of translating the full FLORES-200 *devtest* English set into the language. We estimated GPT model costs using the prompts employed in our experiments, the tiktoken tokenizer[12] used by both models, and inference prices posted by OpenAI.[13] Conveniently, Google Translate costs nothing for the first 500K input characters. But since frequent MT users may have already expended this allowance, we calculated costs from their rates beyond the first 500K.[14] As the NLLB-MOE model (54.5B parameters) is difficult to run on standard computing devices, NLLB Team et al. (2022) also provided a version with only 3.3B parameters that achieves similar performance. Since users commonly opt for the smaller model, and since the performance difference does not impact our estimates significantly, we estimated the costs to run the 3.3B-parameter NLLB model using a single GPU on Google Colab. Details of our estimation method are in Appendix B.1. Table 10 contains the average cost for each system across the languages we evaluated with it.

model	cost
NLLB	$0.09
ChatGPT (0-shot)	$0.35
ChatGPT (5-shot)	$1.32
Google	$2.66
GPT-4 (5-shot)	$25.93

Table 10: Estimated cost in USD to translate FLORES-200 *devtest* ENG→X with each system, averaged across all languages we evaluated with each

Figure 3 displays chrF scores for the 11 languages on which we evaluated all four MT sys-

[12] https://github.com/openai/tiktoken
[13] https://openai.com/pricing
[14] https://cloud.google.com/translate/pricing

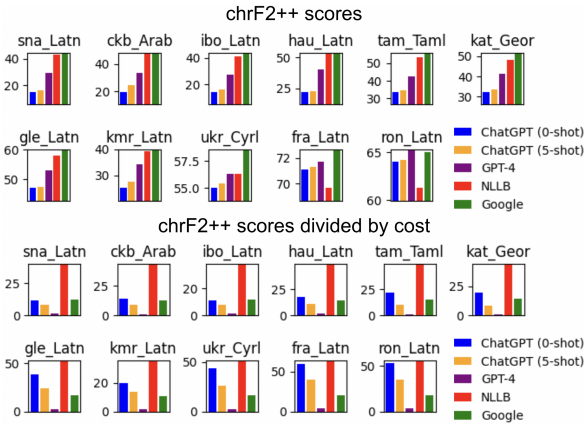


Figure 3: chrF scores for the 11 languages on which we evaluted all MT systems (top), followed by the same scores divided by the estimated cost of each system for each language (bottom)

tems (top), and the same scores divided by the approximate cost of each model (bottom). Bars for GPT-4 drop significantly in the bottom chart because of its high cost. Note from the top chart that Google Translate scores the best, but the bottom chart shows that NLLB has the best scores for its price. Zero-shot ChatGPT also tops five-shot in the bottom chart, suggesting that while few-shot prompts provide modest score improvements, they may not be worth the extra cost. See Appendix B for fuller visualizations with all 203 languages.

4 Related Work

We are not the first researchers to explore LLM MT. However, most existing studies do not provide benchmarks for a large number or languages. Wang et al. (2023) studied GPT model discourse MT, but only for four languages. Gao et al. (2023) studied prompt engineering for GPT model MT, a helpful precursor to our work, but only for three languages. Moslem et al. (2023) probed the abilities of GPT models for adaptive and domain-appropriate MT and term extraction, only including six languages in five directions. Jiao et al. (2023) produced MT benchmarks for ChatGPT and GPT-4, but only for five languages, none of them LRLs.[15] They corroborated our findings that GPT models lag behind traditional MT models, but that GPT-4 outperforms ChatGPT. Hendy et al. (2023) explored 18 language pairs in a similar study, including four LRLs, but they focused more on MT performance across text domains, in-context learning, and reasoning than on multilingual benchmarks.

In all the heretofore mentioned works combined, researchers explored only 18 languages, including five LRLs. This few-language approach does not address the needs of LLM users seeking to translate any languages other than the small few represented. In a work most comparable to our own, Zhu et al. (2023) attempted to address this issue. They provided benchmarks comparing LLMs and traditional MT models across 102 languages, including 68 LRLs. Their results corroborate our own conclusions that LLMs lag behind traditional MT models, especially for LRLs. However, their analysis focuses primarily on few-shot learning and prompt engineering, including some topics somewhat removed from end user needs (such as the viability of nonsensical prompts in few-shot settings). Our work differs from existing studies in our focus on end users. We include more languages than any existing work (**204** languages, including **168** LRLs), to address the needs of various LRL communities. Our analysis suggests which language features predict LLM effectiveness, to help end users make hypotheses even about languages not represented in our study. We evaluate monetary costs, since they are a concern for LLM users.

5 Conclusion

We provide benchmarks for LLM ENG→X MT performance across 203 languages, with comparisons to state-of-the-art commercial and open-source MT models. For many HRLs, LLMs like ChatGPT perform competitively with these traditional models. But for LRLs, traditional MT remains dominant, despite LLMs' increased parameter size. Our decision-tree analysis reveals language features that predict ChatGPT's translation effectiveness relative to NLLB, finding that ChatGPT is especially disadvantaged for LRLs and African languages, and that the number of Wikipedia pages a language has is a strong predictor of ChatGPT's effectiveness in it. We present evidence that few-shot learning offers generally marginal improvements for ENG→X MT, which may not justify its additional cost. We provide MT users with scores and cost estimates for four LLM and traditional MT systems, to help them determine which to use for their languages.

Future work may include more translation directions (X→ENG and non-English-centric), document-level MT, and human evaluation of LLM

[15] In this section, we define LRLs as languages having fewer than 1M Wikipedia pages.

outputs to reveal trends along fluency and accuracy dimensions. We open-source software and outputs of the models we evaluated on our repository.

Limitations

We acknowledge limitations of using ChatGPT models for research. Since they are closed-source models, there is much we do not know about their architectural and training details, which can impact our understanding of their capabilities and biases. For instance, OpenAI's implementation of mechanisms to prevent the generation of harmful or toxic content may inadvertently impact the quality of the model's output. This can be a concern when evaluating the reliability and accuracy of the results. OpenAI continuously updates and deprecates models behind the ChatGPT API, so our assessment may not be immaculate for future versions. Future work may mitigate these concerns by evaluating white-box LLMs, such as BLOOM (Scao et al., 2022) or MPT (Team, 2023), or LLMs not tuned for instruction, like GPT-3 (Brown et al., 2020).

While FLORES-200 is large and diverse, it is likely not representative of the vast array of languages worldwide. Some low-resource sets within FLORES-200 may contain noisy or corrupted data, potentially affecting the validity of the automatic metrics we employ in our reporting of scores. Additionally, FLORES-200 sets were translated from English Wikipedia. We avoided any X→ENG translation directions, since it is likely that GPT models were trained on English Wikipedia. However, the semantic proximity of the other language sets to the original English source could potentially provide an advantage to these models in generating them. We also acknowledge the absence of non-English-centric translation directions from this study; we leave this for future work.

Lastly, the unavailability of semantic MT evaluation techniques like COMET (Rei et al., 2020) or BLEURT (Sellam et al., 2020) for LRLs hinders our ability to conduct comprehensive semantic evaluations and may leave some aspects of the translation quality unexplored. Future researchers may gain additional insights by evaluating LLM COMET scores for the target languages in which they are available. Human evaluation (which we leave for future work) may also reveal much in this area. These limitations surrounding model transparency, representative data, and evaluation should be taken into account when interpreting the findings of this work. Future studies may benefit from addressing these challenges to enhance the robustness and reliability of MT conclusions.

Ethics Statement

The new prominence of LLMs in language technologies has numerous ethical implications. This study makes it apparent that even powerful LLMs like ChatGPT have significant limitations, such as an inability to translate a large number of low-resource languages. It also suggests that although these LLMs are trained on large and diverse data sets, they still have implicit biases, such as a clear disadvantage in MT for African languages. We hope to stress the importance of acknowledging and publicizing the limits and biases of these LLMs. This is especially relevant because a majority of LLM users may not be familiar or experienced with artificial intelligence (AI) engineering practices, and the commercial entities providing LLMs often have a monetary incentive to deliberately downplay the models' limitations. This can lead to unethical exploitation of users, who may attempt to use LLMs in applications where their limitations and biases can cause harm. Part of our goal in this work is to bring these discussions to the forefront of AI research. Ethical considerations like these should be a top concern for AI researchers, especially when many recent AI advancements are piloted by powerful commercial corporations.

We hope also to acknowledge some of the ethical considerations involved in our own research. As we strive to develop improved open-source and accessible translation systems, it is essential to acknowledge that some language communities may have reservations about having their languages translated. Another crucial point is that utilizing the FLORES-200 test set in this research may inadvertently contribute to its incorporation into OpenAI's training data. OpenAI's current position is that API requests are not used for training (Schade, 2023), but if this position were altered or disregarded, it could compromise the reliability of this test set for future GPT iterations. (This is a consideration for many commercial LLMs, though we only used OpenAI's in the current work.) This scenario has a potential negative impact on the MT community, since many researchers depend on FLORES-200 and other MT benchmarks for large, diverse, high-quality data to conduct system comparisons.

Acknowledgements

We thank Simran Khanuja for her help in running our Google Translate baseline and her general support. We also thank Alex Cabrera for his help developing our Zeno browser. This material is based on research sponsored in part by the Air Force Research Laboratory under agreement number FA8750-19-2-0200. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory or the U.S. Government. This work was also supported in part by the National Science Foundation under grant #2040926, a grant from the Singapore Defence Science and Technology Agency.

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. *arXiv e-prints*, page arXiv:2201.06642.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R. Mortensen, Noah A. Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2022. Building machine translation systems for the next thousand languages.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *CoRR*, abs/2010.11125.

Yuan Gao, Ruili Wang, and Feng Hou. 2023. How to design translation prompts for chatgpt: An empirical study. *arXiv e-prints*, pages arXiv–2304.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Wenxiang Jiao, WX Wang, JT Huang, Xing Wang, and ZP Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yasmin Moslem, Rejwanul Haque, and Andy Way. 2023. Adaptive machine translation with large language models. *arXiv preprint arXiv:2301.13294*.

Graham Neubig and Zhiwei He. 2023. Zeno GPT Machine Translation Report.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers,

Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

OpenAI. 2023. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Michael Schade. 2023. How your data is used to improve model performance.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Aditya Siddhant, Ankur Bapna, Orhan Firat, Yuan Cao, Mia Xu Chen, Isaac Caswell, and Xavier Garcia. 2022. Towards the next 1000 languages in multilingual machine translation: Exploring the synergy between supervised and self-supervised learning. *CoRR*, abs/2201.03110.

MosaicML NLP Team. 2023. Introducing mpt-7b: A new standard for open-source, commercially usable llms. Accessed: 2023-05-05.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. *arXiv preprint arXiv:2304.02210*.

Liang Xu, Anqi Li, Lei Zhu, Hang Xue, Changtai Zhu, Kangkang Zhao, Haonan He, Xuanwei Zhang, Qiyue Kang, and Zhenzhong Lan. 2023. Superclue: A comprehensive chinese large language model benchmark.

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis.

A Unabridged Result Table

In Table 11 we report full results for 203 target languages in ENG→X translation directions, across four MT systems: two LLMs (ChatGPT and GPT-4, with two n-shot settings for ChatGPT), one open-source encoder-decoder MT model (NLLB), and one commercial system (Google). We order in them in increasing order of performance, with zero-shot ChatGPT performing the worst and Google performing the best overall. We obtained scores for 203 target languages with ChatGPT, 201 with NLLB, 115 with Google Translate, and 20 with GPT-4. Our scores are spBLEU (Goyal et al., 2022) using the SPM-200 tokenizer (NLLB Team et al., 2022) and chrF2++ (Popović, 2017). All results are also available on our repository, and interactive visualizations and histograms can be browsed on our Zeno browser.

B Unabridged Bar Charts and Cost Estimation

See Figures 4 and 5 for chrF and BLEU scores across all MT systems and languages. Google Translate and NLLB are generally the best performers in both metrics, though GPT-4 and ChatGPT are occasionally best. An "x" indicates where we did not evaluate one of the systems for a language. Figures 6 and 7 display chrF and BLEU scores divided by the estimated cost of each MT system. The cost value is measured as the amount in USD that it would cost to translate the entire FLORES-200 *devtest* set for each language.

These visualizations are also available on our repository. (Also see our Zeno browser for interactive visualizations of our results.) We also include cost estimates and scores divided thereby for all languages and MT systems in Table 14. We exclude cost estimates by language for NLLB and Google because there is very little variation between languages. Our estimated cost of translating FLORES-200 *devtest* ENG→ is approximately $0.09 for every target language. And the respective estimate for Google Translate is roughly $2.66 regardless of the target language, since Google's API only charges for input characters.

B.1 Details about estimating NLLB cost

To estimate the cost of running NLLB's 3.3B-parameter model for translation, we used one GPU from Google Colab to translate the full FLORES-200 *devtest* set from English into six languages representing six high- and low-resource scripts–Burmese (`mya_Mymr`), Simplified Chinese (`zho_Hans`), Standard Arabic (`arb_Arab`), Hindi (`hin_Deva`), Armenian (`hye_Armn`), and French (`fra_Latn`)–and measured the time for each. We assumed that runtime t is determined by an equation with unknown coefficients x_1, x_2, and x_3:

$$t = x_1 n_{input} + x_2 n_{output} + x_3 \quad (1)$$

where n_{input} represents the number of input tokens and n_{output} is the number of output tokens. In this case, x_1 represents the rate at which the encoder processes input tokens, x_2 represents the rate at which the decoder undergoes inference, and x_3 is the amount of time to perform all other computations, independent of the number of tokens. We estimated x_1, x_2, and x_3 via a least-squares solution to the linear system defined by the six languages for which we obtained runtime t:

$$\begin{bmatrix} n_{input} & n_{output}(\texttt{mya}) & 1 \\ n_{input} & n_{output}(\texttt{zho}) & 1 \\ n_{input} & n_{output}(\texttt{arb}) & 1 \\ n_{input} & n_{output}(\texttt{hin}) & 1 \\ n_{input} & n_{output}(\texttt{hye}) & 1 \\ n_{input} & n_{output}(\texttt{fra}) & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} t_{\texttt{mya}} \\ t_{\texttt{zho}} \\ t_{\texttt{arb}} \\ t_{\texttt{hin}} \\ t_{\texttt{hye}} \\ t_{\texttt{fra}} \end{bmatrix}$$

where n_{input} is the number of tokens in the English *devtest* set, and n_{output} for each language is the number of tokens in the NLLB-MOE model output provided by NLLB Team et al. (2022). (We used the same tokenizer that we had used for GPT model cost estimation, for simplicity.) After estimating x_1, x_2, and x_3, we used them in Equation 1 to estimate t values for all 201 languages for which we obtained NLLB MT scores. We then used Google Colab's estimated rate of $0.35/hour for use of one GPU to estimate costs for each language.

C Visualizations Comparing ChatGPT and NLLB

See Figures 8 and 9. They are also posted on our repository. (Also see our Zeno browser for interactive visualizations of our results.)

D Estimating Wikipedia Page Counts

As mentioned in §2.3, we used the "Total pages" count from `https://en.wikipedia.org/wiki/List_of_Wikipedias`, accessed 7 August 2023,

as a proxy for the resource level of a language (refered to as `wiki_ct` in §3.4). We had to make some decisions regarding macrolanguage and microlanguage matches when making these estimates. Many of the languages in FLORES-200 (NLLB Team et al., 2022) are in fact microlanguages of a macrolangauge not included in the dataset. In some cases this microlanguage was did not have a listed Wikipedia page count, so we used the macrolanguage page count instead. Table 12 lists all the languages for which we used the Wikipedia page count of a macrolanguage (with a different ISO 639-3 code), based on our best judgment. In every case this was because the FLORES-200 microlanguage was not listed.

There were also cases where we decided to list zero for a microlanguage's `wiki_ct`, even if its macrolanguage was listed with a nonzero number of pages. This was in cases where we could reasonably assume that the macrolanguage's Wikipedia pages were likely (either all or predominantly) in another microlanguage or dialect. We list the languages that we considered in this manner in Table 13.

We also made some decisions regarding `wiki_ct` assignment based on the script of a language. We recorded zero Wikipedia pages for `kas_Deva` and 13,210 for `kas_Arab` (all of the Kashmiri pages) because a majority of Kashmiri pages seem to be in Perso-Arabic script. (There may be a few in Devanagari, but we simplify by assuming none are.) We also recorded zero pages for `mni_Beng` because, although Wikipedia has pages in Meitei, they appear to be in the Meitei `Mtei` script, not Bengali `Beng`. Lastly, we assigned Wikipedia's count for 'Classical Chinese' (`zh-classical`) to `zho_Hant` and its count for 'Chinese' to `zho_Hans` (though it is possible that some of the 'Chinese' pages may be in the Traditional Chinese (`Hant`) script).

In all other cases, if a language did not have a listed number of Wikipedia pages, we took this to mean it had zero.

Table 11: BLEU and chrF results on ENG→X directions. “0-shot" and “5-shot" are ChatGPT with zero- and five-shot settings, respectively. “NLLB" is the NLLB-MOE model, and “Google" is Google Translate. We used five-shot settings only for GPT-4. Models are listed in order of their effectiveness in MT (with zero-shot ChatGPT performing the worst and Google Translate performing the best).

Language	**spBLEU200**					**chrF2++**				
	0-shot	5-shot	GPT-4	NLLB	Google	0-shot	5-shot	GPT-4	NLLB	Google
ace_Arab	1.3	2.3	–	5.5	–	8.4	9.8	–	17.4	–
ace_Latn	5.0	4.3	–	11.6	–	19.8	18.0	–	37.1	–
acm_Arab	28.2	29.6	29.5	11.8	–	44.7	46.1	46.5	31.9	–
acq_Arab	30.9	31.9	–	26.9	–	47.5	48.1	–	42.2	–
aeb_Arab	24.2	24.7	–	19.9	–	41.0	41.3	–	38.2	–
afr_Latn	47.2	46.7	–	44.4	48.7	67.0	66.7	–	64.3	67.8
ajp_Arab	31.5	32.2	32.2	36.3	–	47.1	47.5	48.4	51.3	–
aka_Latn	3.2	3.1	–	11.7	–	13.3	13.8	–	34.5	–
als_Latn	33.6	34.2	–	39.4	–	56.0	56.3	–	58.3	–
amh_Ethi	3.5	3.7	–	31.6	34.1	10.0	10.6	–	39.4	42.0
apc_Arab	30.4	30.9	–	36.7	–	45.5	45.8	–	50.6	–
arb_Arab	37.6	37.9	–	43.0	48.6	53.8	53.8	–	57.1	62.6
arb_Latn	5.3	8.4	–	–	7.9	22.8	26.9	–	–	35.4
ars_Arab	35.9	37.2	–	36.7	–	52.4	53.1	–	50.5	–
ary_Arab	19.3	19.6	–	23.3	–	36.3	36.7	–	38.9	–
arz_Arab	26.2	26.6	–	32.1	–	42.3	42.7	–	46.8	–
asm_Beng	8.2	10.6	–	22.5	23.2	23.2	26.1	–	35.9	37.4
ast_Latn	31.3	32.3	–	34.5	–	53.8	54.5	–	56.8	–
awa_Deva	15.6	16.6	–	27.6	–	35.4	36.3	–	47.1	–
ayr_Latn	0.2	0.1	–	7.6	7.2	4.7	3.8	–	29.7	31.5
azb_Arab	3.5	3.6	–	5.4	–	17.9	18.5	–	23.5	–
azj_Latn	16.6	17.7	–	24.6	–	38.4	40.3	–	42.9	–
bak_Cyrl	5.5	5.7	–	30.3	–	20.1	20.7	–	47.3	–
bam_Latn	0.5	0.7	–	9.3	9.5	6.1	6.9	–	30.5	32.6
ban_Latn	10.9	9.0	–	19.4	–	30.7	27.4	–	44.6	–
bel_Cyrl	19.5	20.5	–	27.3	30.1	38.3	39.1	–	42.0	44.4
bem_Latn	1.6	1.1	–	13.6	–	10.3	9.1	–	37.9	–
ben_Beng	21.8	22.1	–	36.0	37.6	38.5	39.0	–	50.0	51.4
bho_Deva	11.9	12.5	–	23.6	21.0	29.7	30.7	–	42.8	40.0
bjn_Arab	2.0	3.0	–	5.8	–	10.4	13.2	–	17.1	–
bjn_Latn	11.0	12.3	–	21.9	–	35.9	38.0	–	48.2	–
bod_Tibt	0.2	0.4	–	8.5	–	12.7	14.7	–	29.7	–
bos_Latn	40.0	40.6	–	40.7	44.0	59.9	60.1	–	58.8	61.8
bug_Latn	5.2	2.7	–	9.1	–	23.3	16.4	–	33.7	–
bul_Cyrl	44.1	44.4	–	50.0	53.1	61.6	61.9	–	64.8	67.9
cat_Latn	47.8	47.9	–	48.9	51.1	65.4	65.3	–	65.0	67.2
ceb_Latn	28.0	29.1	–	34.5	40.2	51.0	52.9	–	57.3	62.2
ces_Latn	40.8	40.8	–	42.4	46.0	57.6	57.4	–	57.4	60.3
cjk_Latn	0.2	0.1	–	4.0	–	4.4	4.5	–	24.3	–
ckb_Arab	4.7	6.5	11.2	26.8	25.8	19.7	24.8	33.1	47.2	47.7
crh_Latn	6.0	6.8	–	27.4	–	27.8	29.0	–	47.0	–
cym_Latn	48.0	48.5	–	58.4	63.6	64.7	64.9	–	70.8	74.5
dan_Latn	52.3	52.5	–	50.0	55.3	69.7	69.7	–	66.4	70.3
deu_Latn	47.7	47.9	–	46.6	51.2	65.4	65.4	–	62.8	66.5
dik_Latn	0.2	0.1	–	6.1	–	4.6	4.4	–	24.2	–
dyu_Latn	0.5	0.1	–	2.7	–	4.5	4.3	–	17.7	–
dzo_Tibt	0.1	0.7	–	13.3	–	7.7	15.9	–	34.7	–
ell_Grek	35.8	35.8	–	38.7	40.1	51.6	51.6	–	52.0	53.6
epo_Latn	37.9	38.5	–	42.8	40.4	58.5	58.8	–	61.4	60.1
est_Latn	35.3	35.8	–	36.5	41.4	56.8	56.9	–	56.1	59.9
eus_Latn	19.1	19.5	–	29.0	33.9	44.2	43.9	–	50.0	54.5
ewe_Latn	0.6	0.7	–	17.2	17.0	6.0	6.1	–	39.0	39.9
fao_Latn	18.1	19.2	–	31.6	–	40.5	41.5	–	49.8	–
fij_Latn	5.5	4.8	–	23.6	–	22.9	21.3	–	46.7	–
fin_Latn	35.8	36.1	–	36.6	39.2	56.2	56.4	–	55.3	58.0
fon_Latn	0.2	0.2	–	6.4	–	3.9	4.1	–	21.5	–
fra_Latn	56.4	56.6	57.3	56.2	59.7	71.1	71.3	71.7	69.7	72.7
fur_Latn	18.5	19.8	–	39.6	–	40.6	42.5	–	56.8	–
fuv_Latn	1.2	0.4	–	6.0	–	8.5	5.8	–	23.9	–
gaz_Latn	0.6	0.4	–	12.6	14.6	8.0	7.3	–	37.5	40.3
gla_Latn	15.5	16.3	–	28.7	32.2	38.9	39.0	–	50.2	52.7

Continued on next page

Table 11 – continued from previous page

Language	spBLEU200					chrF2++				
	0-shot	5-shot	GPT-4	NLLB	Google	0-shot	5-shot	GPT-4	NLLB	Google
gle_Latn	25.8	26.3	32.8	41.4	44.1	47.1	47.5	53.0	58.0	60.1
glg_Latn	39.4	40.0	–	40.1	41.9	61.3	61.5	–	59.8	61.5
grn_Latn	0.7	0.6	–	16.4	15.3	6.3	5.7	–	36.6	36.4
guj_Gujr	18.9	19.4	–	37.2	39.2	37.4	37.1	–	53.3	55.2
hat_Latn	24.5	24.8	–	30.5	31.8	47.0	47.2	–	51.9	53.4
hau_Latn	6.2	6.3	15.7	31.4	30.6	22.2	22.4	40.3	53.5	53.2
heb_Hebr	35.3	35.4	–	46.8	48.8	51.2	50.7	–	59.8	61.2
hin_Deva	29.2	29.4	–	40.6	43.0	48.7	48.6	–	57.3	59.3
hne_Deva	14.1	15.5	–	33.7	–	34.0	36.1	–	54.3	–
hrv_Latn	37.8	38.2	–	38.9	42.5	57.0	57.2	–	57.2	60.2
hun_Latn	34.8	34.9	–	38.1	40.9	54.6	54.5	–	55.5	58.1
hye_Armn	14.3	14.8	–	40.2	42.7	33.2	33.5	–	53.2	56.3
ibo_Latn	3.2	4.0	9.8	20.6	22.2	14.7	16.3	27.7	41.4	43.5
ilo_Latn	11.4	12.6	–	29.0	31.0	33.6	35.6	–	53.3	56.0
ind_Latn	48.8	48.7	–	49.2	55.0	68.5	68.5	–	68.7	72.6
isl_Latn	26.0	26.0	–	33.9	40.8	44.8	45.0	–	50.0	55.8
ita_Latn	37.6	37.7	–	38.3	40.0	59.4	59.5	–	57.3	59.1
jav_Latn	16.9	18.9	–	30.3	30.3	41.2	42.7	–	54.8	55.1
jpn_Jpan	30.5	31.3	–	20.1	35.3	33.1	33.7	–	27.9	37.1
kab_Latn	1.3	1.5	–	16.9	–	11.9	12.9	–	35.6	–
kac_Latn	0.0	0.1	–	14.3	–	2.9	4.8	–	37.5	–
kam_Latn	1.3	1.1	–	6.1	–	8.9	9.0	–	25.9	–
kan_Knda	18.6	19.4	–	39.6	41.9	37.9	38.2	–	53.4	55.7
kas_Arab	4.0	3.3	–	18.2	–	15.5	14.3	–	34.2	–
kas_Deva	2.3	2.7	–	4.7	–	12.9	13.9	–	17.1	–
kat_Geor	15.2	15.7	23.2	34.6	37.5	32.5	33.5	41.4	48.1	51.4
kaz_Cyrl	12.9	13.4	–	34.0	38.7	33.9	33.4	–	51.8	56.0
kbp_Latn	0.4	1.4	–	11.3	–	4.0	9.4	–	28.3	–
kea_Latn	13.0	18.7	–	22.5	–	37.6	43.0	–	42.8	–
khk_Cyrl	8.0	8.5	–	27.1	33.1	26.1	26.6	–	43.9	49.8
khm_Khmr	5.7	6.0	–	23.0	27.4	21.5	21.1	–	36.4	40.3
kik_Latn	0.8	2.0	–	15.4	–	8.8	11.6	–	37.1	–
kin_Latn	3.4	3.1	–	27.2	34.3	18.7	18.0	–	49.7	56.1
kir_Cyrl	8.4	8.9	–	27.4	30.5	25.8	26.6	–	44.5	48.2
kmb_Latn	0.4	0.4	–	4.5	–	4.9	6.1	–	24.9	–
kmr_Latn	8.3	9.4	14.3	19.6	20.0	25.3	27.4	34.3	39.3	40.0
knc_Arab	0.5	1.1	–	6.5	–	5.3	4.7	–	9.8	–
knc_Latn	2.6	0.9	–	8.2	–	13.4	8.1	–	27.4	–
kon_Latn	0.9	1.3	–	18.9	–	8.5	10.5	–	45.3	–
kor_Hang	25.6	25.9	–	26.7	30.0	34.4	34.9	–	36.0	38.6
lao_Laoo	2.9	4.0	–	29.6	29.6	18.5	21.5	–	46.2	44.0
lij_Latn	7.6	10.3	–	37.2	–	32.8	35.2	–	53.8	–
lim_Latn	15.1	19.8	21.0	25.8	–	40.2	42.7	45.1	47.9	–
lin_Latn	2.6	2.5	–	21.9	21.4	14.8	14.7	–	48.0	48.4
lit_Latn	30.0	30.6	–	35.4	41.7	51.5	51.8	–	54.7	59.4
lmo_Latn	6.7	8.3	–	10.5	–	29.9	30.6	–	34.9	–
ltg_Latn	5.3	5.4	–	36.4	–	29.2	29.1	–	53.6	–
ltz_Latn	25.4	27.5	–	36.7	35.3	48.7	48.9	–	56.0	55.6
lua_Latn	1.0	1.1	–	9.8	–	8.1	9.3	–	35.2	–
lug_Latn	1.6	1.3	–	14.0	14.4	11.6	10.6	–	39.8	41.3
luo_Latn	0.8	0.1	–	15.2	–	7.0	5.0	–	38.5	–
lus_Latn	4.6	4.7	–	15.1	–	17.6	17.8	–	38.0	–
lvs_Latn	33.0	33.5	36.7	35.4	–	55.1	55.2	57.3	54.8	–
mag_Deva	18.6	19.4	24.8	39.4	–	39.1	39.9	44.6	58.5	–
mai_Deva	10.2	12.1	–	27.1	19.6	28.9	31.2	–	46.7	40.6
mal_Mlym	14.6	14.9	–	38.3	43.2	32.3	32.0	–	51.6	56.2
mar_Deva	14.5	14.7	–	30.3	33.4	34.3	34.6	–	48.0	51.0
min_Arab	1.6	3.5	–	–	–	10.1	14.9	–	–	–
min_Latn	11.5	13.1	–	28.7	–	37.0	38.4	–	52.4	–
mkd_Cyrl	36.0	36.5	–	42.6	46.5	57.0	57.3	–	60.6	63.7
mlt_Latn	29.9	30.3	–	50.3	59.7	49.4	49.8	–	66.0	71.6
mni_Beng	1.8	2.0	–	27.5	0.1	11.4	10.5	–	38.7	0.6
mos_Latn	0.2	0.2	–	6.8	–	3.9	4.3	–	24.3	–
mri_Latn	15.1	14.5	–	20.7	18.3	34.8	34.0	–	44.2	42.4
mya_Mymr	2.1	2.8	–	17.7	24.5	19.8	20.6	–	32.0	40.4
nld_Latn	36.3	36.5	–	35.6	38.0	56.5	56.7	–	54.9	57.3

Continued on next page

Table 11 – continued from previous page

Language	spBLEU200					chrF2++				
	0-shot	5-shot	GPT-4	NLLB	Google	0-shot	5-shot	GPT-4	NLLB	Google
nno_Latn	37.1	38.3	–	33.4	25.6	58.7	59.4	–	53.6	50.7
nob_Latn	40.2	39.8	–	38.4	–	60.5	60.2	–	58.6	–
npi_Deva	19.0	19.6	–	28.7	–	39.5	39.3	–	45.5	–
nso_Latn	3.7	4.6	–	26.5	29.8	16.7	19.0	–	50.8	54.0
nus_Latn	0.1	0.5	–	14.4	–	3.0	5.5	–	29.0	–
nya_Latn	4.9	5.5	–	17.7	21.1	20.6	22.6	–	44.0	48.0
oci_Latn	30.4	33.3	–	41.0	–	55.1	57.0	–	58.8	–
ory_Orya	11.6	12.6	–	30.2	38.9	27.5	29.8	–	45.7	53.4
pag_Latn	5.7	8.3	–	20.2	–	22.6	26.7	–	46.3	–
pan_Guru	21.0	21.5	–	36.4	39.7	37.4	37.6	–	49.0	51.9
pap_Latn	25.4	33.2	–	42.2	–	51.6	56.5	–	60.2	–
pbt_Arab	5.1	5.8	9.2	22.9	–	19.7	21.1	26.7	39.4	–
pes_Arab	29.4	30.4	–	36.1	39.8	48.6	48.8	–	51.3	54.3
plt_Latn	8.2	8.3	–	25.3	25.9	31.4	30.9	–	50.0	51.2
pol_Latn	32.1	32.6	–	32.5	36.3	49.7	50.0	–	48.9	52.1
por_Latn	56.4	56.9	–	52.9	58.6	71.4	71.7	–	67.9	72.3
prs_Arab	25.7	27.5	–	33.8	–	44.8	47.4	–	53.6	–
quy_Latn	0.7	0.6	–	5.8	8.2	9.3	9.5	–	26.9	34.0
ron_Latn	46.2	46.9	49.0	44.7	50.0	64.0	64.2	65.3	61.3	65.0
run_Latn	3.1	2.3	–	19.6	–	16.6	14.7	–	42.5	–
rus_Cyrl	38.9	38.9	–	41.0	43.9	56.6	56.5	–	56.3	58.7
sag_Latn	0.1	0.1	–	10.5	–	4.6	5.1	–	35.7	–
san_Deva	4.7	5.4	–	8.0	10.0	21.8	22.6	–	26.1	30.3
sat_Olck	0.0	1.9	–	18.5	–	0.2	14.4	–	26.3	–
scn_Latn	11.2	13.0	–	24.4	–	35.9	37.2	–	46.8	–
shn_Mymr	0.5	1.3	–	15.1	–	7.6	16.6	–	34.4	–
sin_Sinh	6.1	6.9	–	36.0	40.4	19.5	20.1	–	43.8	51.2
slk_Latn	38.6	38.4	–	42.9	48.4	56.8	57.0	–	59.0	63.1
slv_Latn	35.7	36.0	–	38.1	42.4	55.5	55.7	–	56.2	59.6
smo_Latn	6.3	8.0	–	26.9	–	22.8	26.3	–	50.0	–
sna_Latn	3.2	3.4	8.4	19.7	20.8	15.3	16.3	29.2	43.4	44.4
snd_Arab	9.1	10.5	–	31.9	32.6	22.5	24.9	–	48.1	48.7
som_Latn	8.1	8.1	–	18.4	18.9	29.4	29.7	–	43.0	43.7
sot_Latn	5.7	5.4	–	20.7	22.5	20.7	20.9	–	46.1	47.8
spa_Latn	33.8	33.9	–	33.1	35.0	56.5	56.7	–	53.8	55.5
srd_Latn	16.3	18.5	–	35.8	–	42.1	43.8	–	55.6	–
srp_Cyrl	37.5	37.9	–	43.4	48.1	56.5	57.2	–	59.7	63.4
ssw_Latn	1.9	0.5	5.8	19.9	–	10.6	6.7	24.1	43.3	–
sun_Latn	13.9	14.5	–	21.6	24.4	39.0	38.6	–	44.7	48.7
swe_Latn	52.5	52.2	–	50.1	54.2	68.5	68.4	–	65.9	69.4
swh_Latn	38.0	38.6	–	36.8	44.6	60.1	60.3	–	58.6	64.4
szl_Latn	12.8	15.1	–	38.4	–	35.5	36.7	–	53.7	–
tam_Taml	13.6	13.4	20.9	36.6	38.7	33.8	34.5	42.7	53.7	55.8
taq_Latn	0.8	0.3	–	4.9	–	8.2	6.2	–	23.1	–
taq_Tfng	0.6	1.4	–	5.6	–	5.2	8.3	–	16.7	–
tat_Cyrl	6.7	7.3	–	30.4	30.4	21.5	23.6	–	46.8	48.2
tel_Telu	17.4	18.0	–	41.6	44.7	34.4	35.6	–	55.9	58.2
tgk_Cyrl	10.8	11.7	–	35.3	35.6	29.3	30.4	–	51.2	51.8
tgl_Latn	35.0	35.0	–	38.3	39.8	60.8	60.6	–	60.5	61.8
tha_Thai	33.5	33.6	–	35.1	45.2	43.1	43.2	–	42.7	49.7
tir_Ethi	1.6	1.9	–	17.8	17.6	5.8	6.7	–	25.8	26.3
tpi_Latn	14.0	15.8	22.7	17.8	–	37.1	39.2	49.5	41.6	–
tsn_Latn	3.8	4.2	–	25.6	–	17.0	18.6	–	48.5	–
tso_Latn	2.8	3.0	–	26.7	26.1	15.0	16.0	–	50.0	50.9
tuk_Latn	6.2	7.7	–	22.6	35.8	25.2	25.9	–	42.1	52.7
tum_Latn	3.6	2.9	–	13.3	–	16.5	14.8	–	35.2	–
tur_Latn	38.5	38.5	–	41.5	46.4	57.9	57.8	–	58.3	62.4
twi_Latn	3.0	3.0	–	15.2	17.4	13.4	14.2	–	37.9	40.9
tzm_Tfng	1.1	2.2	–	21.0	–	8.3	11.7	–	32.3	–
uig_Arab	6.5	8.5	–	30.5	40.2	20.5	24.7	–	45.3	54.3
ukr_Cyrl	37.4	37.4	39.2	40.1	42.8	55.0	55.4	56.3	56.3	58.6
umb_Latn	0.4	0.1	–	4.1	–	5.3	4.9	–	26.6	–
urd_Arab	21.9	22.2	–	30.5	32.7	41.7	41.8	–	48.9	50.0
uzn_Latn	17.4	18.8	–	30.0	37.8	39.9	40.9	–	50.6	56.4
vec_Latn	15.7	17.5	–	28.2	–	41.0	42.8	–	51.6	–
vie_Latn	40.7	40.7	–	43.3	–	58.5	57.9	–	59.5	–

Continued on next page

Table 11 – continued from previous page

Language	spBLEU200					chrF2++				
	0-shot	5-shot	GPT-4	NLLB	Google	0-shot	5-shot	GPT-4	NLLB	Google
war_Latn	24.3	25.0	28.4	35.0	–	49.3	49.5	54.0	57.4	–
wol_Latn	2.1	1.2	–	9.6	–	10.6	8.3	–	29.7	–
xho_Latn	5.3	6.0	–	25.4	29.5	21.9	23.3	–	48.6	52.2
ydd_Hebr	10.6	18.7	–	18.4	16.8	31.0	38.1	–	38.6	37.7
yor_Latn	2.5	3.3	–	10.5	4.9	11.4	13.7	–	25.5	20.0
yue_Hant	26.4	33.8	–	16.6	–	22.3	27.2	–	17.9	–
zho_Hans	36.3	36.5	–	26.6	43.6	31.0	31.9	–	22.8	37.8
zho_Hant	29.3	30.4	–	12.4	–	24.8	26.0	–	14.0	–
zsm_Latn	41.4	41.3	–	45.5	47.5	64.5	64.3	–	66.5	68.0
zul_Latn	6.7	7.3	–	31.4	32.0	25.2	26.3	–	53.3	53.9

Figure 4: chrF scores across all MT systems and languages

Figure 5: BLEU scores across all MT systems and languages

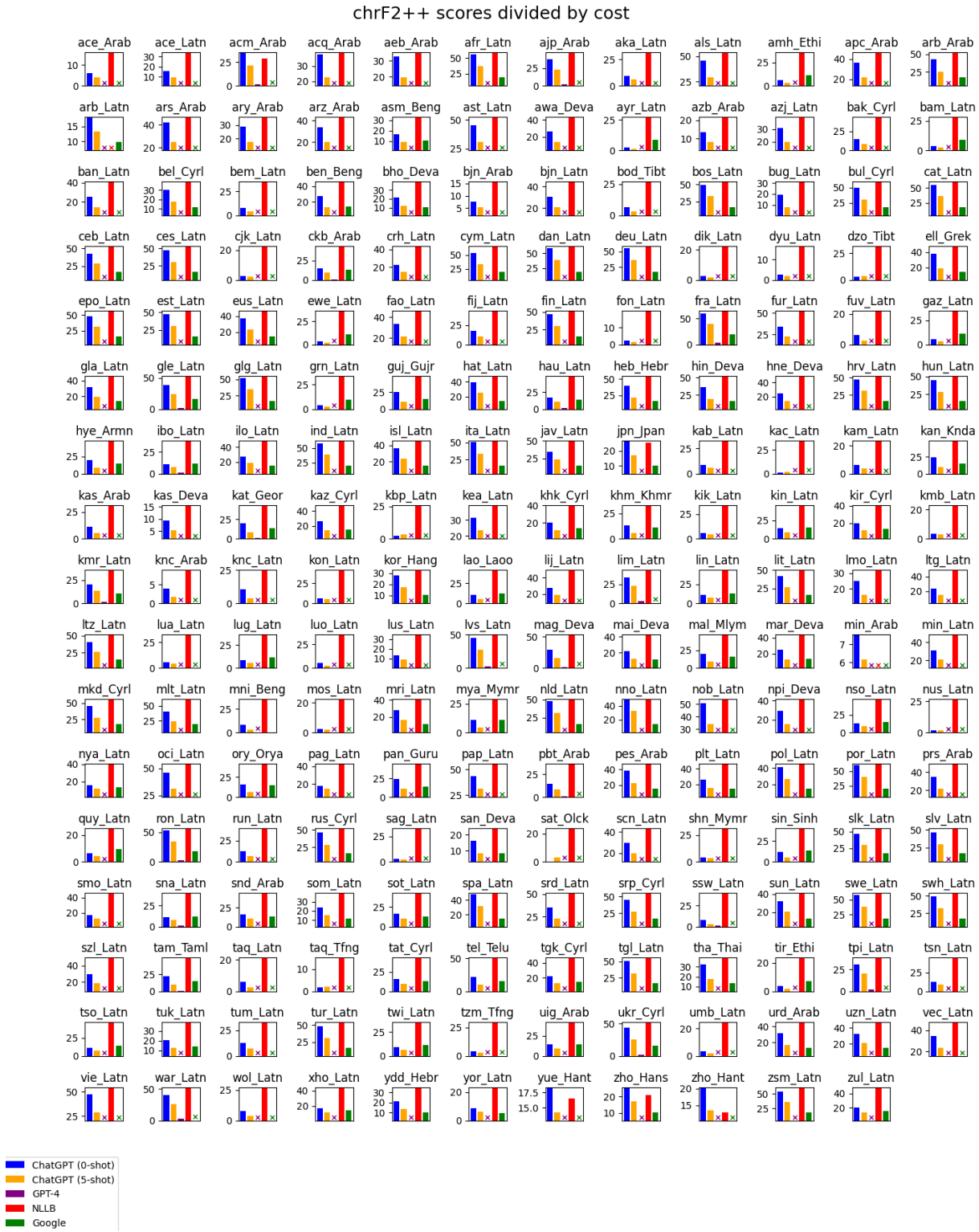


Figure 6: chrF scores divided by the estimated cost of each MT system, across all MT systems and languages

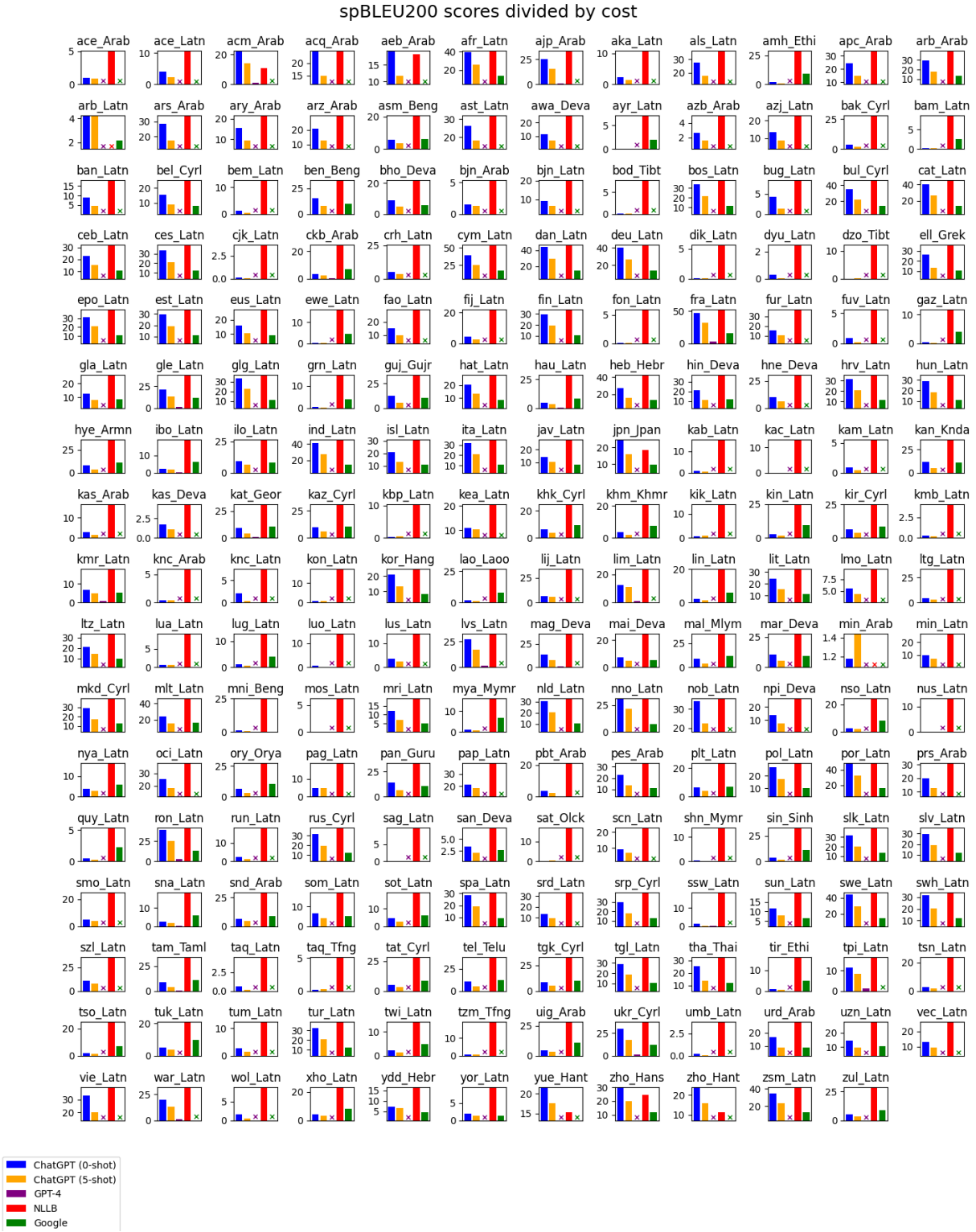


Figure 7: BLEU scores divided by the estimated cost of each MT system, across all MT systems and languages

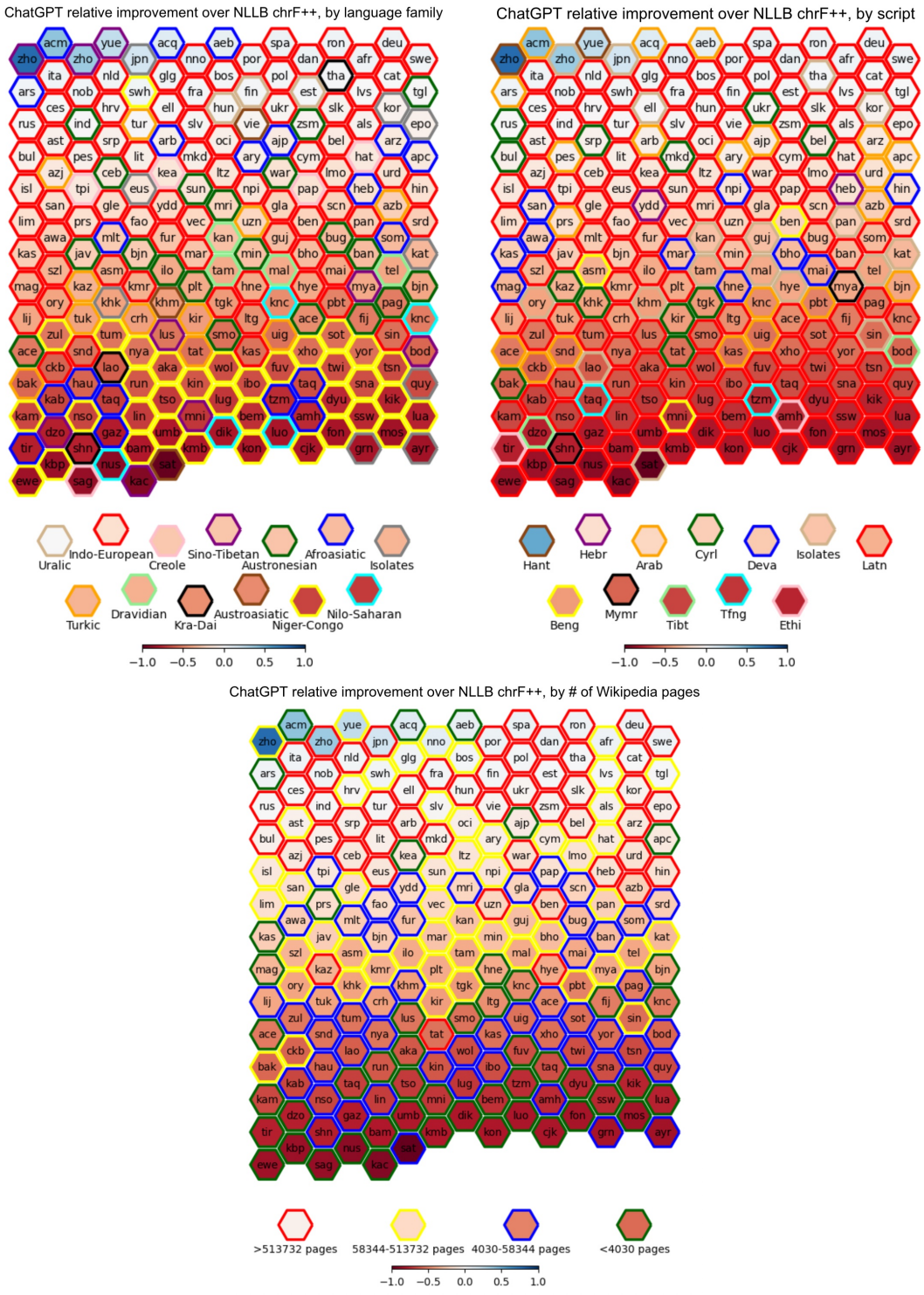


Figure 8: ChatGPT *relative improvement* over NLLB chrF (color scale), with languages organized by family, script, and number of Wikipedia pages (divided in quartiles). Hexagons (one per language) are displayed in descending order across rows, with the highest ChatGPT relative improvement over NLLB chrF2++ at the top left, and the lowest at the bottom right. Group hexagons at the bottom of each plot display the average color for each group and are organized in like manner.

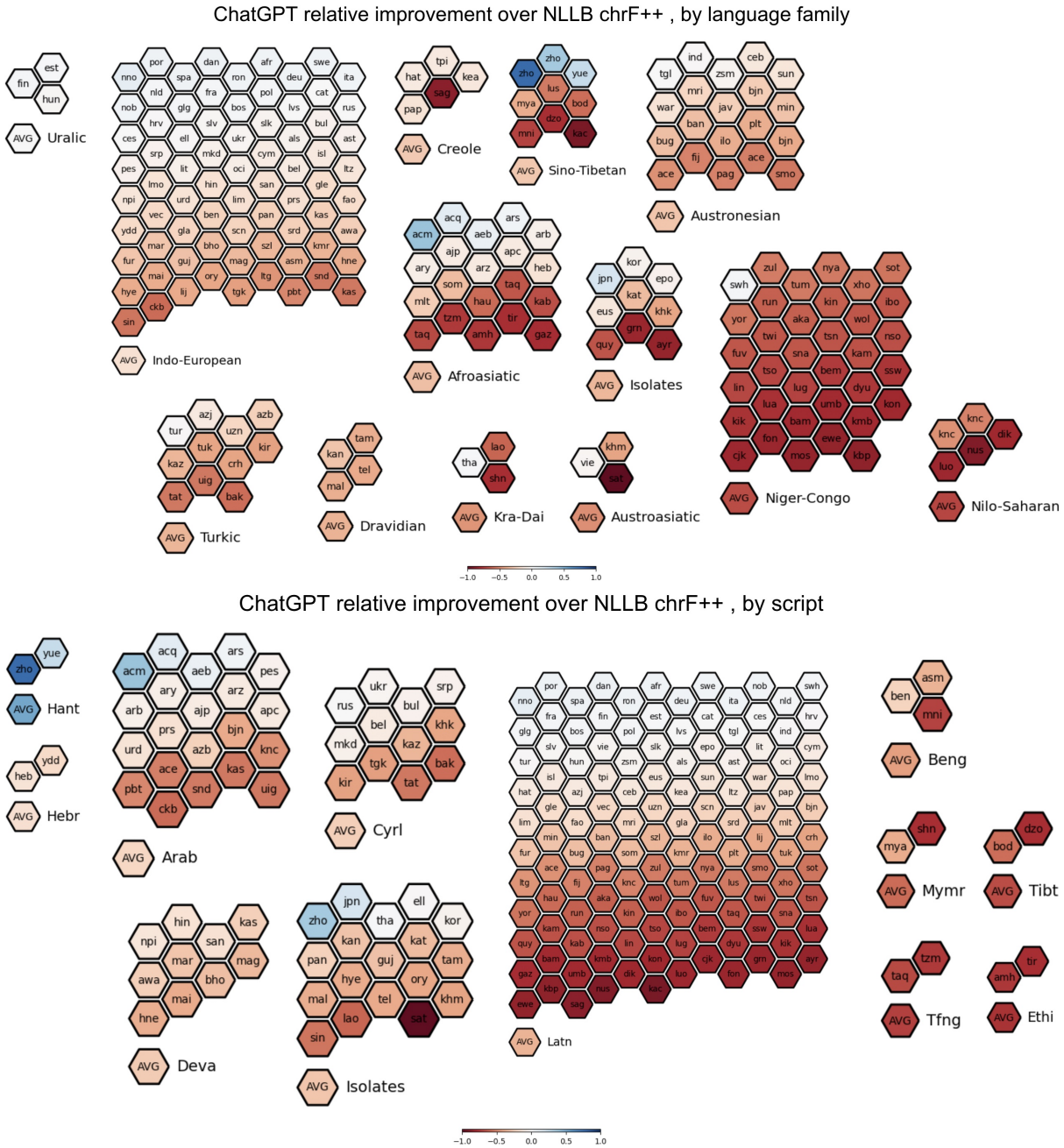


Figure 9: Alternative visualizations to those in Figure 8. Groups and languages are organized the same here: from top left to bottom right in descending order of the ChatGPT *relative improvement* over NLLB (using averages for the groups).

FLORES lang.	substitution for wiki_ct
arb	Used macrolanguage ‘Arabic’ (ara) because ‘Standard Arabic’ (arb) not present
bho	Used macrolanguage ‘Bihari’ (bih) because ‘Bhojpuri’ (bho) not present
dik	Used macrolanguage ‘Dinka’ (din) because ‘Southwestern Dinka’ (dik) not present
fuv	Used macrolanguage ‘Fula’ (ful) because "Nigerian Fulfulde" (fuv) not present
knc	Used macrolanguage ‘Kanuri’ (kau) because ‘Central Kanuri’ (knc) not present
lvs	Used macrolanguage ‘Latvian’ (lav) because ‘Standard Latvian’ (lvs) not present
plt	Used macrolanguage ‘Malagasy’ (mlg) because ‘Plateau Malagasy’ (plt) not present
khk	Used macrolanguage ‘Mongolian’ (mon) because ‘Halh Mongolian’ (khk) not present
gaz	Used macrolanguage ‘Oromo’ (orm) because ‘West Central Oromo’ (gaz) not present
pes	Used macrolanguage ‘Persian’ (fas) because ‘Western Persian’ (pes) not present
pbt	Used macrolanguage ‘Pashto’ (pus) because ‘Southern Pashto’ (pbt) not present
quy	Used macrolanguage ‘Quechua’ (que) because ‘Ayuacucho Quechua’ (quy) not present
als	Used macrolanguage ‘Albanian’ (sqi) because ‘Tosk Albanian’ (als) not present
uzn	Used macrolanguage ‘Uzbek’ (uzb) because ‘Northern Uzbek’ (uzn) not present
ydd	Used macrolanguage ‘Yiddish’ (yid) because ‘Eastern Yiddish’ (ydd) not present
zsm	Used macrolangauge ‘Malay’ (msa) because ‘Standard Malay’ (zsm) not present

Table 12: FLORES-200 languages for which we used the Wikipedia page count associated with a macrolanguage of another ISO 639-3 code

FLORES lang.	reason for assigning wiki_ct = 0
acm	Macrolanguage ‘Arabic’ (ara) appears to be in ‘Standard Arabic’ (arb), not ‘Mesopotamian Arabic’ (acm)
acq	Macrolanguage ‘Arabic’ (ara) appears to be in ‘Standard Arabic’ (arb), not ‘Tai’izzi Arabic’ (acq)
aeb	Macrolanguage ‘Arabic’ (ara) appears to be in ‘Standard Arabic’ (arb), not ‘Tunisian Arabic’ (aeb)
ajp	Macrolanguage ‘Arabic’ (ara) appears to be in ‘Standard Arabic’ (arb), not ‘South Levantine Arabic’ (ajp)
apc	Macrolanguage ‘Arabic’ (ara) appears to be in ‘Standard Arabic’ (arb), not ‘North Levantine Arabic’ (apc)
ars	Macrolanguage ‘Arabic’ (ara) appears to be in ‘Standard Arabic’ (arb), not ‘Najdi Arabic’ (ars)
mag	Macrolanguage ‘Bihari’ (bih) appears to be in ‘Bhojpuri’ (bho), not ‘Magahi’ (mag)
prs	Macrolanguage ‘Persian’ (fas) appears to be in ‘Western Persian’ (pes), not ‘Dari’ (prs)

Table 13: FLORES-200 languages for which we used assigned wiki_ct to be zero, despite the existence of Wikipedia pages in a corresponding macrolanguage

Table 14: Estimated costs in USD to translate the FLORES-200 *devtest* set ENG→X for each targe language and MT system, along with BLEU and chrF scores divided by the cost estimates, where applicable. The cost is roughly $0.09 for NLLB and $2.66 for Google Translate for all target languages.

Lang.	spBLEU200/cost					chrF2++/cost					cost estimate (USD$)		
	0-shot	5-shot	GPT-4	NLLB	Google	0-shot	5-shot	GPT-4	NLLB	Google	0-shot	5-shot	GPT-4
ace_Arab	0.9	0.9	–	5.1	–	6.3	4.0	–	16.0	–	0.3	1.5	29.0
ace_Latn	4.0	2.2	–	10.7	–	15.8	9.1	–	34.2	–	0.3	1.0	18.9
acm_Arab	22.4	13.9	1.2	10.8	–	35.6	21.6	1.8	29.3	–	0.3	1.1	24.3
acq_Arab	24.6	14.9	–	24.8	–	37.7	22.4	–	38.9	–	0.3	1.1	24.6
aeb_Arab	19.3	11.6	–	18.3	–	32.7	19.4	–	35.2	–	0.3	1.1	24.1
afr_Latn	39.5	25.8	–	41.0	13.3	56.2	36.8	–	59.4	18.5	0.2	0.8	17.1
ajp_Arab	25.0	15.3	1.3	33.4	–	37.4	22.5	2.0	47.3	–	0.3	1.1	23.7
aka_Latn	2.3	1.4	–	10.8	–	9.5	6.1	–	31.8	–	0.4	1.2	22.9
als_Latn	27.6	17.5	–	36.3	–	46.0	28.9	–	53.8	–	0.2	0.9	20.3
amh_Ethi	2.3	1.1	–	28.8	9.3	6.4	3.2	–	35.9	11.5	0.6	2.4	50.8
apc_Arab	24.2	14.7	–	33.8	–	36.2	21.8	–	46.6	–	0.3	1.1	23.6
arb_Arab	29.9	17.6	–	39.6	13.3	42.7	25.0	–	52.6	17.1	0.3	1.1	24.8
arb_Latn	4.2	4.2	–	–	2.1	18.1	13.4	–	–	9.7	0.3	1.0	21.5
ars_Arab	28.5	17.3	–	33.8	–	41.7	24.7	–	46.5	–	0.3	1.2	24.8
ary_Arab	15.3	9.2	–	21.5	–	28.8	17.2	–	35.8	–	0.3	1.1	24.4
arz_Arab	20.9	12.5	–	29.6	–	33.7	20.0	–	43.1	–	0.3	1.1	24.3
asm_Beng	5.7	3.6	–	20.6	6.4	16.1	8.8	–	32.9	10.2	0.4	2.0	42.6
ast_Latn	26.4	18.1	–	31.9	–	45.3	30.6	–	52.5	–	0.2	0.8	16.5
awa_Deva	11.5	6.4	–	25.3	–	26.1	14.0	–	43.2	–	0.4	1.6	34.6
ayr_Latn	0.1	0.0	–	7.0	2.0	2.8	1.5	–	27.4	8.6	0.7	1.5	19.9
azb_Arab	2.7	1.6	–	5.0	–	13.6	8.1	–	21.6	–	0.3	1.3	26.4
azj_Latn	13.4	8.7	–	22.7	–	31.1	19.7	–	39.6	–	0.2	1.0	22.5
bak_Cyrl	4.0	2.2	–	27.9	–	14.8	8.2	–	43.5	–	0.4	1.5	31.8
bam_Latn	0.3	0.3	–	8.6	2.6	3.7	2.8	–	28.1	8.9	0.6	1.4	22.1
ban_Latn	9.0	4.8	–	17.9	–	25.4	14.6	–	41.2	–	0.2	0.9	17.7
bel_Cyrl	15.3	9.1	–	25.1	8.2	30.0	17.3	–	38.7	12.1	0.3	1.3	27.4
bem_Latn	1.1	0.5	–	12.5	–	7.3	4.1	–	35.0	–	0.4	1.2	20.1
ben_Beng	15.4	7.7	–	33.0	10.3	27.2	13.6	–	45.8	14.0	0.4	1.9	40.6
bho_Deva	8.8	4.8	–	21.7	5.7	21.9	11.9	–	39.3	10.9	0.4	1.6	34.2
bjn_Arab	1.5	1.3	–	5.3	–	7.7	5.5	–	15.7	–	0.4	1.4	29.0
bjn_Latn	9.2	6.8	–	20.2	–	30.2	21.0	–	44.5	–	0.2	0.8	17.2
bod_Tibt	0.1	0.1	–	7.7	–	6.4	3.3	–	26.9	–	1.0	3.4	71.3
bos_Latn	33.3	21.9	–	37.6	12.0	49.9	32.4	–	54.3	16.9	0.2	0.9	18.2

Continued on next page

Table 14 – continued from previous page

Lang.	spBLEU200/cost					chrF2++/cost					cost estimate (USD$)		
	0-shot	5-shot	GPT-4	NLLB	Google	0-shot	5-shot	GPT-4	NLLB	Google	0-shot	5-shot	GPT-4
bug_Latn	4.3	1.4	–	8.4	–	19.2	8.4	–	31.1	–	0.2	1.0	18.7
bul_Cyrl	35.4	21.6	–	46.1	14.5	49.5	30.2	–	59.7	18.5	0.2	1.1	22.5
cat_Latn	40.0	26.4	–	45.2	14.0	54.8	36.0	–	60.0	18.3	0.2	0.8	17.2
ceb_Latn	23.1	15.5	–	31.8	11.0	42.1	28.2	–	52.9	17.0	0.2	0.9	18.5
ces_Latn	33.7	21.3	–	39.1	12.6	47.5	30.0	–	53.0	16.5	0.2	0.9	19.5
cjk_Latn	0.1	0.1	–	3.7	–	2.8	2.0	–	22.4	–	0.6	1.3	18.7
ckb_Arab	3.5	2.5	0.3	24.6	7.0	14.5	9.4	0.9	43.3	13.0	0.4	1.6	35.1
crh_Latn	4.9	3.6	–	25.3	–	23.0	15.2	–	43.4	–	0.2	0.9	19.5
cym_Latn	39.6	25.2	–	53.9	17.4	53.3	33.8	–	65.3	20.4	0.2	0.9	19.7
dan_Latn	44.0	29.3	–	46.2	15.1	58.7	38.9	–	61.3	19.2	0.2	0.8	16.8
deu_Latn	40.2	26.9	–	43.0	14.0	55.1	36.7	–	58.0	18.2	0.2	0.8	16.5
dik_Latn	0.1	0.1	–	5.6	–	2.8	1.8	–	22.3	–	0.6	1.5	20.3
dyu_Latn	0.3	0.1	–	2.5	–	2.7	1.8	–	16.3	–	0.7	1.4	19.2
dzo_Tibt	0.0	0.1	–	12.0	–	2.8	3.5	–	31.4	–	1.8	3.6	76.8
ell_Grek	26.1	13.3	–	35.5	10.9	37.7	19.2	–	47.7	14.6	0.4	1.7	36.8
epo_Latn	31.5	20.8	–	39.5	11.0	48.7	31.7	–	56.7	16.4	0.2	0.9	18.1
est_Latn	29.4	19.3	–	33.7	11.3	47.3	30.7	–	51.8	16.4	0.2	0.9	18.2
eus_Latn	15.9	10.5	–	26.8	9.3	36.8	23.6	–	46.2	14.9	0.2	0.9	18.2
ewe_Latn	0.4	0.3	–	15.8	4.6	3.7	2.4	–	35.9	10.9	0.6	1.6	23.1
fao_Latn	15.0	10.1	–	29.2	–	33.4	21.8	–	45.9	–	0.2	0.9	19.3
fij_Latn	4.3	2.4	–	21.8	–	17.8	10.5	–	43.1	–	0.3	1.0	19.8
fin_Latn	29.6	19.1	–	33.8	10.7	46.6	29.9	–	51.0	15.8	0.2	0.9	18.9
fon_Latn	0.1	0.1	–	5.9	–	2.3	1.4	–	19.8	–	0.7	1.9	28.4
fra_Latn	47.5	31.7	3.3	51.9	16.3	60.0	39.9	4.1	64.4	19.9	0.2	0.8	16.6
fur_Latn	15.4	10.7	–	36.6	–	33.8	23.0	–	52.4	–	0.2	0.9	18.0
fuv_Latn	0.9	0.2	–	5.5	–	6.2	2.6	–	22.0	–	0.4	1.2	18.0
gaz_Latn	0.4	0.2	–	11.6	4.0	5.5	3.3	–	34.6	11.0	0.5	1.2	20.7
gla_Latn	12.5	8.1	–	26.5	8.8	31.5	19.5	–	46.3	14.4	0.2	1.0	21.2
gle_Latn	21.1	13.3	1.5	38.2	12.0	38.5	24.1	2.4	53.5	16.4	0.2	1.0	20.7
glg_Latn	33.3	22.6	–	37.0	11.5	51.8	34.7	–	55.2	16.8	0.2	0.8	16.4
grn_Latn	0.5	0.2	–	15.1	4.2	4.1	2.4	–	33.8	9.9	0.5	1.3	19.8
guj_Gujr	12.5	5.8	–	33.9	10.7	24.7	11.1	–	48.6	15.1	0.5	2.4	51.2
hat_Latn	20.5	13.6	–	28.2	8.7	39.3	25.8	–	47.9	14.6	0.2	0.8	17.4
hau_Latn	4.9	3.2	0.8	29.0	8.4	17.6	11.5	2.0	49.4	14.5	0.3	0.9	18.9
heb_Hebr	27.2	15.3	–	43.0	13.3	39.4	21.9	–	55.0	16.7	0.3	1.3	28.3
hin_Deva	21.5	11.3	–	37.3	11.8	35.8	18.7	–	52.6	16.2	0.4	1.6	34.6
hne_Deva	10.4	6.0	–	30.9	–	25.0	14.0	–	49.8	–	0.4	1.6	34.1
hrv_Latn	31.5	20.7	–	35.9	11.6	47.5	30.9	–	52.8	16.5	0.2	0.8	18.0
hun_Latn	28.6	18.1	–	35.1	11.2	44.9	28.3	–	51.2	15.9	0.2	0.9	19.7
hye_Armn	8.6	3.7	–	36.5	11.7	19.9	8.5	–	48.4	15.4	0.7	2.9	63.9
ibo_Latn	2.4	1.9	0.4	19.0	6.1	11.0	7.8	1.2	38.2	11.9	0.3	1.1	21.9
ilo_Latn	9.3	6.5	–	26.8	8.5	27.3	18.5	–	49.2	15.3	0.2	0.9	19.2
ind_Latn	41.1	27.5	–	45.4	15.0	57.8	38.6	–	63.5	19.8	0.2	0.8	16.3
isl_Latn	21.3	13.5	–	31.3	11.2	36.8	23.4	–	46.1	15.3	0.2	0.9	19.7
ita_Latn	31.7	21.0	–	35.4	10.9	50.1	33.2	–	52.9	16.1	0.2	0.8	16.8
jav_Latn	14.1	10.3	–	28.0	8.3	34.4	23.4	–	50.6	15.0	0.2	0.8	17.3
jpn_Jpan	24.9	16.0	–	18.5	9.7	27.1	17.2	–	25.7	10.1	0.2	1.0	20.4
kab_Latn	1.0	0.7	–	15.6	–	9.1	6.2	–	32.8	–	0.3	1.1	21.5
kac_Latn	0.0	0.0	–	13.2	–	1.6	1.9	–	34.6	–	0.8	1.5	21.1
kam_Latn	0.9	0.5	–	5.6	–	6.4	4.1	–	23.9	–	0.4	1.2	19.9
kan_Knda	11.8	5.3	–	36.0	11.5	24.1	10.5	–	48.6	15.2	0.6	2.6	58.0
kas_Arab	2.9	1.2	–	16.7	–	11.3	5.4	–	31.4	–	0.4	1.7	33.8
kas_Deva	1.7	1.0	–	4.3	–	9.3	5.4	–	15.7	–	0.4	1.6	34.1
kat_Geor	9.3	4.0	0.4	31.5	10.2	19.8	8.6	0.6	43.8	14.1	0.6	2.9	62.7
kaz_Cyrl	9.8	5.7	–	31.3	10.6	25.9	14.1	–	47.6	15.3	0.3	1.4	29.1
kbp_Latn	0.2	0.5	–	10.4	–	2.3	3.3	–	26.0	–	0.7	1.9	34.3
kea_Latn	11.0	10.3	–	20.8	–	31.6	23.6	–	39.5	–	0.2	0.8	17.3
khk_Cyrl	6.1	3.6	–	24.9	9.1	19.9	11.3	–	40.4	13.6	0.3	1.4	28.9
khm_Khmr	3.5	1.7	–	21.0	7.5	13.3	5.8	–	33.2	11.0	0.6	2.6	57.2
kik_Latn	0.5	0.8	–	14.2	–	6.0	4.6	–	34.2	–	0.5	1.5	26.1
kin_Latn	2.6	1.5	–	25.1	9.4	14.6	9.0	–	45.8	15.3	0.3	1.0	19.6
kir_Cyrl	6.4	3.9	–	25.2	8.3	19.6	11.5	–	41.0	13.2	0.3	1.3	27.5
kmb_Latn	0.3	0.2	–	4.2	–	3.2	2.7	–	23.0	–	0.5	1.3	19.7
kmr_Latn	6.6	4.8	0.7	18.1	5.5	20.2	14.0	1.6	36.3	10.9	0.3	1.0	20.1
knc_Arab	0.4	0.4	–	6.0	–	3.9	1.8	–	9.0	–	0.3	1.6	28.2
knc_Latn	2.1	0.4	–	7.6	–	10.6	3.6	–	25.3	–	0.3	1.2	21.1
kon_Latn	0.7	0.6	–	17.4	–	5.9	5.0	–	41.8	–	0.4	1.1	18.8
kor_Hang	20.9	13.1	–	24.6	8.2	28.1	17.6	–	33.2	10.5	0.2	1.0	21.0
lao_Laoo	1.6	1.0	–	26.9	8.1	10.2	5.5	–	42.0	12.0	0.8	2.9	62.0
lij_Latn	6.3	5.5	–	34.3	–	27.4	18.8	–	49.6	–	0.2	0.9	18.8
lim_Latn	12.6	10.8	1.1	23.8	–	33.6	23.2	2.4	44.2	–	0.2	0.8	17.8
lin_Latn	2.0	1.3	–	20.2	5.9	11.2	7.4	–	44.3	13.2	0.3	1.0	18.2
lit_Latn	24.6	15.8	–	32.7	11.4	42.2	26.7	–	50.5	16.2	0.2	0.9	20.0
lmo_Latn	5.6	4.4	–	9.7	–	25.0	16.1	–	32.2	–	0.2	0.9	19.1
ltg_Latn	4.4	2.8	–	33.6	–	23.9	15.0	–	49.4	–	0.2	0.9	20.0
ltz_Latn	21.1	14.6	–	33.9	9.7	40.4	25.9	–	51.7	15.2	0.2	0.9	18.8
lua_Latn	0.7	0.5	–	9.0	–	5.7	4.4	–	32.5	–	0.4	1.1	18.6
lug_Latn	1.2	0.6	–	12.9	3.9	8.6	5.1	–	36.7	11.3	0.4	1.1	18.7
luo_Latn	0.5	0.1	–	14.0	–	5.0	2.2	–	35.5	–	0.4	1.2	17.9
lus_Latn	3.5	2.4	–	13.9	–	13.5	9.0	–	35.1	–	0.3	1.0	18.8
lvs_Latn	27.0	16.9	1.7	32.7	–	45.1	27.9	2.6	50.5	–	0.2	1.0	20.7
mag_Deva	13.7	7.5	0.7	36.2	–	28.8	15.4	1.3	53.7	–	0.4	1.6	34.3
mai_Deva	7.5	4.6	–	24.9	5.4	21.3	11.9	–	42.9	11.1	0.4	1.6	35.3

Continued on next page

Table 14 – continued from previous page

Lang.	spBLEU200/cost					chrF2++/cost					cost estimate (USD$)		
	0-shot	5-shot	GPT-4	NLLB	Google	0-shot	5-shot	GPT-4	NLLB	Google	0-shot	5-shot	GPT-4
mal_Mlym	9.1	4.0	–	34.9	11.8	20.1	8.7	–	47.0	15.4	0.6	2.7	58.5
mar_Deva	10.5	5.5	–	27.8	9.1	24.8	12.9	–	44.0	13.9	0.4	1.7	36.3
min_Arab	1.2	1.4	–	–	–	7.6	6.2	–	–	–	0.3	1.4	30.1
min_Latn	9.7	7.1	–	26.5	–	31.1	21.0	–	48.4	–	0.2	0.8	17.6
mkd_Cyrl	28.9	17.5	–	39.3	12.7	45.6	27.5	–	55.8	17.4	0.2	1.1	23.2
mlt_Latn	24.2	15.1	–	46.4	16.3	40.0	24.9	–	60.8	19.6	0.2	1.0	21.3
mni_Beng	1.3	0.6	–	25.1	0.0	8.0	3.3	–	35.4	0.2	0.4	2.2	45.8
mos_Latn	0.1	0.1	–	6.3	–	2.3	1.8	–	22.4	–	0.7	1.4	20.5
mri_Latn	12.0	7.2	–	19.1	5.0	27.8	16.9	–	40.8	11.6	0.3	1.0	21.0
mya_Mymr	1.2	0.6	–	16.1	6.7	10.9	4.7	–	29.1	11.0	0.8	3.4	73.8
nld_Latn	30.5	20.4	–	32.9	10.4	47.6	31.7	–	50.7	15.7	0.2	0.8	16.6
nno_Latn	31.3	21.4	–	30.8	7.0	49.4	33.1	–	49.5	13.8	0.2	0.8	16.8
nob_Latn	33.9	22.4	–	35.5	–	51.0	33.9	–	54.1	–	0.2	0.8	16.4
npi_Deva	13.9	7.5	–	26.4	–	29.0	15.0	–	41.8	–	0.4	1.6	34.7
nso_Latn	2.8	2.3	–	24.4	8.1	12.5	9.3	–	46.8	14.8	0.3	1.0	19.9
nus_Latn	0.1	0.2	–	13.2	–	1.8	1.9	–	26.6	–	0.7	1.9	30.3
nya_Latn	3.8	2.8	–	16.3	5.8	16.1	11.5	–	40.6	13.1	0.3	1.0	19.4
oci_Latn	25.4	18.1	–	37.9	–	46.1	31.0	–	54.3	–	0.2	0.8	17.9
ory_Orya	6.5	2.8	–	27.3	10.6	15.4	6.6	–	41.4	14.6	0.8	3.5	78.3
pag_Latn	4.6	4.6	–	18.6	–	18.2	14.7	–	42.7	–	0.2	0.8	16.3
pan_Guru	13.9	6.3	–	33.2	10.9	24.8	11.1	–	44.7	14.2	0.5	2.4	52.3
pap_Latn	21.4	18.2	–	39.0	–	43.5	31.0	–	55.6	–	0.2	0.8	17.4
pbt_Arab	3.8	2.4	0.3	21.1	–	14.8	8.9	0.9	36.2	–	0.3	1.4	29.1
pes_Arab	23.1	13.7	–	33.2	10.9	38.0	22.0	–	47.2	14.8	0.3	1.2	26.1
plt_Latn	6.6	4.2	–	23.3	7.1	25.4	15.6	–	46.1	14.0	0.2	1.0	20.3
pol_Latn	26.7	17.5	–	30.0	9.9	41.3	26.8	–	45.1	14.2	0.2	0.9	18.4
por_Latn	47.8	32.4	–	48.9	16.0	60.5	40.8	–	62.7	19.8	0.2	0.8	15.9
prs_Arab	20.2	12.6	–	31.1	–	35.1	21.6	–	49.3	–	0.3	1.2	25.4
quy_Latn	0.5	0.3	–	5.4	2.2	6.5	4.4	–	24.8	9.3	0.4	1.2	19.4
ron_Latn	38.6	25.3	2.6	41.3	13.7	53.3	34.6	3.4	56.6	17.8	0.2	0.9	18.1
run_Latn	2.4	1.2	–	18.1	–	12.8	7.2	–	39.2	–	0.3	1.0	19.7
rus_Cyrl	31.6	19.4	–	37.8	12.0	46.0	28.2	–	51.9	16.1	0.2	1.0	21.5
sag_Latn	0.1	0.0	–	9.7	–	2.8	2.1	–	32.9	–	0.6	1.4	19.3
san_Deva	3.5	2.0	–	7.3	2.7	16.1	8.5	–	24.0	8.3	0.4	1.7	35.7
sat_Olck	0.0	0.4	–	16.8	–	0.1	3.1	–	23.8	–	1.1	3.6	80.2
scn_Latn	9.3	6.9	–	22.5	–	29.9	19.7	–	43.2	–	0.2	0.9	18.9
shn_Mymr	0.3	0.3	–	13.6	–	4.1	3.2	–	31.0	–	0.8	4.1	93.0
sin_Sinh	3.7	1.9	–	32.8	11.0	11.9	5.5	–	39.9	14.0	0.6	2.7	57.0
slk_Latn	31.8	20.0	–	39.6	13.2	46.8	29.7	–	54.4	17.2	0.2	0.9	19.6
slv_Latn	29.7	19.3	–	35.2	11.6	46.2	30.0	–	51.9	16.3	0.2	0.9	18.2
smo_Latn	4.8	3.9	–	24.8	–	17.5	13.0	–	46.1	–	0.3	1.0	20.5
sna_Latn	2.4	1.7	0.4	18.2	5.7	11.5	8.0	1.4	40.0	12.1	0.3	1.0	20.3
snd_Arab	6.7	4.3	–	29.3	8.9	16.5	10.2	–	44.2	13.3	0.4	1.4	30.2
som_Latn	6.6	4.1	–	17.0	5.2	24.0	15.2	–	39.7	11.9	0.2	1.0	20.1
sot_Latn	4.5	2.7	–	19.1	6.1	16.1	10.4	–	42.5	13.1	0.3	1.0	20.1
spa_Latn	28.6	19.2	–	30.6	9.6	47.9	32.1	–	49.7	15.2	0.2	0.8	16.3
srd_Latn	13.6	9.8	–	33.0	–	35.0	23.2	–	51.3	–	0.2	0.9	18.8
srp_Cyrl	29.9	17.9	–	40.0	13.1	45.1	27.0	–	55.0	17.3	0.3	1.1	24.0
ssw_Latn	1.3	0.2	0.3	18.4	–	7.5	2.9	1.1	40.0	–	0.4	1.3	21.1
sun_Latn	11.6	7.9	–	19.9	6.7	32.5	20.9	–	41.3	13.3	0.2	0.8	17.8
swe_Latn	44.2	29.3	–	46.3	14.8	57.7	38.5	–	60.9	19.0	0.2	0.8	16.5
swh_Latn	31.5	20.5	–	34.0	12.2	49.8	32.1	–	54.1	17.6	0.2	0.9	18.6
szl_Latn	10.6	7.8	–	35.4	–	29.4	19.0	–	49.5	–	0.2	0.9	19.9
tam_Taml	8.8	4.0	0.4	33.4	10.6	22.0	10.3	0.8	49.0	15.3	0.5	2.4	51.2
taq_Latn	0.6	0.1	–	4.5	–	5.8	2.7	–	21.3	–	0.4	1.3	20.1
taq_Tfng	0.2	0.3	–	5.1	–	1.8	2.0	–	15.1	–	1.9	3.2	65.1
tat_Cyrl	5.0	3.1	–	28.0	8.3	16.1	9.9	–	43.1	13.2	0.3	1.4	28.8
tel_Telu	11.1	5.1	–	37.9	12.2	21.9	10.1	–	50.9	15.9	0.6	2.5	54.8
tgk_Cyrl	8.2	5.0	–	32.5	9.7	22.3	13.1	–	47.1	14.1	0.3	1.3	28.1
tgl_Latn	29.1	18.4	–	35.3	10.9	50.5	31.9	–	55.8	16.9	0.2	0.9	19.2
tha_Thai	25.4	13.5	–	32.3	12.4	32.7	17.4	–	39.2	13.6	0.3	1.5	32.3
tir_Ethi	1.0	0.6	–	16.2	4.8	3.6	2.0	–	23.5	7.2	0.6	2.4	51.9
tpi_Latn	11.5	8.3	1.1	16.4	–	30.5	20.6	2.5	38.4	–	0.2	0.9	18.9
tsn_Latn	2.9	2.0	–	23.6	–	12.8	9.0	–	44.7	–	0.3	1.1	20.4
tso_Latn	2.1	1.5	–	24.6	7.1	11.3	7.7	–	46.1	13.9	0.3	1.1	20.4
tuk_Latn	5.0	3.9	–	20.8	9.8	20.5	12.9	–	38.8	14.4	0.2	1.0	21.2
tum_Latn	2.8	1.3	–	12.3	–	12.8	6.9	–	32.5	–	0.3	1.1	22.2
tur_Latn	32.0	20.6	–	38.3	12.7	48.1	31.0	–	53.8	17.1	0.2	0.9	18.4
twi_Latn	2.1	1.4	–	14.0	4.7	9.5	6.5	–	34.9	11.2	0.4	1.2	21.8
tzm_Tfng	0.6	0.5	–	19.0	–	4.3	2.8	–	29.3	–	0.9	3.1	64.7
uig_Arab	4.7	3.1	–	28.0	11.0	14.8	9.1	–	41.5	14.8	0.4	1.7	37.0
ukr_Cyrl	29.6	17.5	1.5	36.9	11.7	43.6	25.9	2.2	51.9	16.0	0.3	1.1	24.4
umb_Latn	0.2	0.1	–	3.8	–	3.6	2.1	–	24.5	–	0.5	1.3	19.1
urd_Arab	16.5	8.9	–	28.0	8.9	31.3	16.8	–	44.9	13.7	0.3	1.5	32.2
uzn_Latn	14.3	9.7	–	27.7	10.3	32.7	21.1	–	46.7	15.4	0.2	0.9	19.8
vec_Latn	13.2	9.6	–	26.0	–	34.4	23.6	–	47.6	–	0.2	0.8	17.2
vie_Latn	33.1	20.4	–	39.9	–	47.6	28.9	–	54.9	–	0.2	1.0	21.4
war_Latn	20.0	13.3	1.4	32.3	–	40.7	26.3	2.7	53.0	–	0.2	0.9	18.7
wol_Latn	1.6	0.6	–	8.9	–	7.9	4.0	–	27.4	–	0.3	1.1	18.5
xho_Latn	4.2	3.1	–	23.4	8.1	17.4	12.0	–	44.8	14.3	0.3	0.9	19.2
ydd_Hebr	7.3	6.6	–	16.8	4.6	21.4	13.5	–	35.3	10.3	0.4	1.8	39.1
yor_Latn	1.9	1.5	–	9.7	1.3	8.7	6.1	–	23.5	5.5	0.3	1.2	24.3
yue_Hant	21.7	17.7	–	15.3	–	18.4	14.2	–	16.5	–	0.2	0.9	19.4

Continued on next page

Table 14 – continued from previous page

Lang.	spBLEU200/cost					chrF2++/cost					cost estimate (USD$)		
	0-shot	5-shot	GPT-4	NLLB	Google	0-shot	5-shot	GPT-4	NLLB	Google	0-shot	5-shot	GPT-4
zho_Hans	30.4	19.7	–	24.6	11.9	25.9	17.2	–	21.1	10.3	0.2	0.9	18.1
zho_Hant	24.1	15.8	–	11.5	–	20.5	13.5	–	12.9	–	0.2	0.9	19.7
zsm_Latn	34.8	23.1	–	42.0	13.0	54.2	35.9	–	61.4	18.6	0.2	0.8	16.7
zul_Latn	5.4	3.7	–	29.0	8.8	20.1	13.3	–	49.2	14.7	0.3	1.0	20.1

Large language models effectively leverage document-level context for literary translation, but critical errors persist

Marzena Karpinska **Mohit Iyyer**
University of Massachusetts Amherst
{mkarpinska, miyyer}@cs.umass.edu
https://litmt.org/

Abstract

Large language models (LLMs) are competitive with the state of the art on a wide range of *sentence-level* translation datasets. However, their ability to translate paragraphs and documents remains unexplored because evaluation in these settings is costly and difficult. We show through a rigorous human evaluation that asking the GPT-3.5 (`text-davinci-003`) LLM to translate an entire *literary* paragraph (e.g., from a novel) at once results in higher-quality translations than standard sentence-by-sentence translation across 18 linguistically-diverse language pairs (e.g., translating into and out of Japanese, Polish, and English). Our evaluation, which took approximately 350 hours of effort for annotation and analysis, is conducted by hiring translators fluent in both the source and target language and asking them to provide both span-level error annotations as well as preference judgments of which system's translations are better. We observe that discourse-level LLM translators commit fewer mistranslations, grammar errors, and stylistic inconsistencies than sentence-level approaches. With that said, critical errors still abound, including occasional content omissions, and a human translator's intervention remains necessary to ensure that the author's voice remains intact. We publicly release our dataset and error annotations to spur future research on the evaluation of document-level literary translation.[1]

1 Introduction

Large language models (LLMs) such as ChatGPT (OpenAI, 2022) demonstrate remarkable performance as stand-alone translation systems, rivaling and sometimes surpassing commercial models on sentence-level benchmarks (Vilar et al., 2022; Hendy et al., 2023; Jiao et al., 2023). Furthermore, LLMs are increasingly being deployed for *document-level* translation (Book Maker, 2023; Pawlak, 2023), a scenario for which there are currently no reliable automatic evaluation methods. In this paper, we hire human translators to conduct a rigorous fine-grained evaluation of GPT-3.5's ability to translate **paragraph-level** texts from **literary works** across 18 different language pairs. Our results (Figure 1) demonstrate that GPT-3.5[2] effectively leverages discourse-level context to produce higher-quality translations than when translating sentences in isolation.

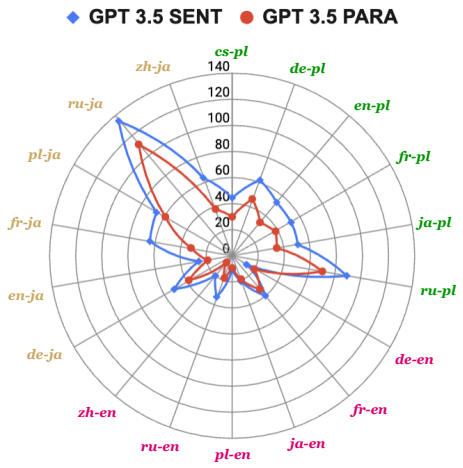


Figure 1: A plot of the total number of errors annotated in sentence-level (SENT) and paragraph-level (PARA) translations produced by GPT-3.5 across 18 different language pairs. In all cases, PARA produces fewer errors than SENT, which demonstrates that GPT-3.5 takes advantage of discourse context during translation.

Why literary texts? Translating works of literature poses unique challenges due to the intricate nature of creative work and the importance of capturing the author's voice and contextual nuances. Translators thus apply a wide range of transla-

[1] https://github.com/marzenakrp/LiteraryTranslation

[2] We completed our annotations on translations from the `text-davinci-003` checkpoint obtained prior to the API release of ChatGPT and GPT-4. Nevertheless, we include a preliminary analysis of GPT-4's translations in §F.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 419–451
December 6–7, 2023.

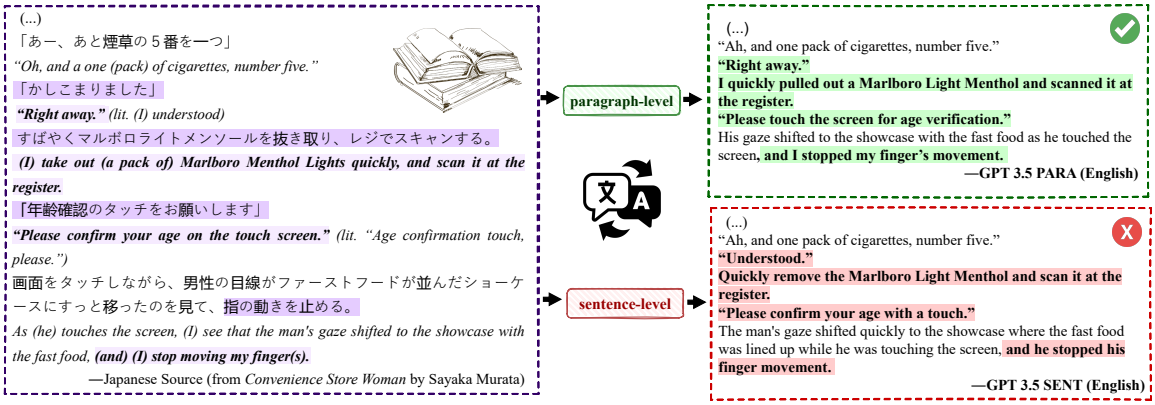


Figure 2: An example of paragraph-level (PARA) and sentence-level (SENT) translations of the same Japanese paragraph into English. Sentence-level translation results in a range of erroneous translations, from worse word choice ("understood" vs "right away") to incorrect pronouns ("he" vs "I"); these errors are corrected by PARA.

tion techniques (Chesterman, 1997; Molina and Hurtado Albir, 2004), from simple shifts in grammatical categories to more complex stylistic or content-based rearrangements that often cross sentence boundaries. Translators may also merge or split sentences and paragraphs, which renders the traditional sentence-level pipeline insufficient for capturing the full scope of the original text (Toral and Way, 2015; Taivalkoski-Shilov, 2019b; Post and Junczys-Dowmunt, 2023; Jiang et al., 2023).[3] Taken together, these properties make literary texts a good testbed for document-level machine translation (Thai et al., 2022); in our work, we focus on the *paragraph*[4] as a minimal discourse-level unit.

Why human evaluation? The absence of rigorous document-level evaluations of LLM translators is striking but also somewhat understandable given the unreliability of automatic metrics (Thai et al., 2022) and the difficulty of properly conducting human evaluations (Castilho, 2021). Furthermore, evaluations of LLM translators are especially difficult due to data contamination (Aiyappa et al., 2023; Chang et al., 2023), as it is unclear whether the models are pretrained on existing benchmarks (e.g., from WMT). We fill this gap by first collecting paragraphs from recently-published literary translations. Then, we provide human translators with two candidate machine translations of a given source paragraph and ask them to (1) mark error *spans* and categorize them based on a predefined schema inspired by MQM (Lommel et al., 2014b; Freitag et al., 2021), (2) make preference judgments of which of the two translations is of higher quality, and (3) provide free-form justifications of their preference judgments. In total, we collect such annotations on **720** pairs of translated paragraphs across **18** different language pairs (using three diverse target languages of English, Japanese, and Polish), which we then leverage for a fine-grained analysis of the behavior of different LLM translation methods.

LLMs produce better translations when provided with paragraph-level context: Our evaluations reveal that using GPT-3.5 to translate complete paragraphs via few-shot prompting (PARA) yields translations of significantly higher quality than both the sentence-by-sentence GPT-3.5 methods (SENT, PARA_SENT) as well as Google Translate. Our detailed analysis of annotated translation errors and free-form comments shows that PARA exhibit increased coherence, better preservation of literary style, and improved handling of context-dependent expressions (see Figure 2). That said, PARA makes many critical mistranslations and other errors across different language pairs, which shows that LLM-based translators still have significant room to improve, particularly when translating contextually-rich literary texts.

2 Background

Before describing our dataset and evaluation, we first contextualize our work within the recent body of research on translation via large language models. We also survey the broader body of document-

[3] At least 55% of the reference target paragraphs used in our study split or merge sentences from the source text (measured with an automatic sentence tokenizer).

[4] We broadly define a paragraph as a distinct passage within the novel, focusing on a single theme.

level[5] MT research in §A.

Translation with large language models: LLM-based translation is attractive because a single model, without training or fine-tuning on *large* parallel corpora, can produce high-quality translations across many language pairs.[6] Recent work explores LLMs' capabilities in this space (Wang et al., 2023) spanning paragraph-level post-editing with LLMs (Thai et al., 2022), translating sentence-level inputs (Vilar et al., 2022; Jiao et al., 2023), analyzing hallucinations in LLM-generated translations (Guerreiro et al., 2023), and employing LLMs to evaluate machine translation (Kocmi and Federmann, 2023). Simple sentence-level English prompt templates have been found effective for paragraph translations (Zhang et al., 2023), and automatically-generated dictionaries can assist LLM-based translation (Ghazvininejad et al., 2023; Lu et al., 2023) along with selecting high-quality demonstrations (Vilar et al., 2022). To the best of our knowledge, the only prior work other than ours that evaluates LLMs for paragraph-level translation is Hendy et al. (2023), who conduct *automatic* evaluation of context-aware sentence-by-sentence translation; in contrast, we perform a fine-grained *human* evaluation of paragraph-level translation.

3 Data & methods

Our work differs from existing research on translating with large language models in two key ways: we focus on translating *literary* text at the *paragraph level*. In this section, we describe and motivate the paragraph-level translation dataset used in our study, which covers 18 unique language pairs (three target languages) and is sourced from recently-published novels. Then, we outline the different ways in which we leverage GPT-3.5 to translate these paragraphs at both the sentence and paragraph levels.

3.1 Dataset collection

Literary texts (e.g., novels or short stories) pose unique challenges for translators due to their complex nature. Translators must interpret and honor the author's voice with no objective reality to measure against, which can result in several equally valid translations (Sager, 1998). For machine translation systems, these challenges exacerbate the need for discourse-level context (Thai et al., 2022): an author's intended meaning or style is often unclear from just a single sentence.

Selecting paragraphs from novels: How good are machines at translating literary paragraphs? To answer this question, we extract **20** paragraphs (dialogues and narrative texts) each from **18** recently-published translations of novels, and we manually align these paragraphs with corresponding paragraphs in the source novel[7] (see Table 8 in §B). Almost all of the translations were published after 2021 (see Table 7 in §B), which is important to avoid data contamination with LLM pretraining data (Aiyappa et al., 2023; Chang et al., 2023). In sum, we obtain **360** aligned source-target paragraphs, which we use for all of the experiments described in the rest of the paper.

Data memorization issue: In order to investigate the extent to which `text-davinci-003` may have memorized the novels in our dataset, we employ the prompts from (Chang et al., 2023) and assess the model's ability to produce masked characters' names. For this purpose we select 171 translation paragraphs, which contained character's names, resulting in an average of 8 out of 20 paragraphs used per book. In nearly all instances, the model was unable to accurately produce the correct names, with three exceptions. Two of these were names of well-known historical figures, "Napoleon Bonaparte" and "Simonides of Ceos." A closer examination revealed that these names could likely be inferred from the context, rather than being a result of the model's memorization. In the third instance the model produced the correct name but in diminutive instead of augmentative form ("Kasia" instead of "Kaśka").[8]

Additionally, we tested `text-davinci-003` with a randomly selected subset of paragraphs from our dataset. In these cases, the model was unable to generate accurate completions.

[5] Note that the term "document-level" has been used in MT research to denote both multi-sentence passages as well as complete documents.

[6] That said, parallel data is almost certainly included in LLM pretraining data, at least for high-resource languages (Briakou et al., 2023).

[7] We purchase the source ebook and its corresponding translation before extracting aligned paragraphs.

[8] Kasia/Kaśka" are both forms of "Katarzyna," the second most common female name in Poland as of January 2023. This raises a question of whether the model's response was due to memorization or an educated guess based on the name's popularity (`https://www.statista.com/statistics/1089014/poland-most-popular-female-names/`).

Paragraph length: All paragraphs consist of at least two sentences, and the majority of them are between four to nine sentences long (mean=7.45, std=4.14).[9] As automatic sentence tokenizers are not always reliable for all of the languages considered in our study, we manually perform sentence tokenization to enable a direct comparison of sentence and paragraph-level translation systems. For more details about the dataset statistics, including token and sentence counts, see §B, which also includes data on sentence numbers obtained using a sentence tokenizer.

Source and target languages: As source languages, we select eight languages that belong to different language families, have varied morphological traits, and employ different writing systems: English (*en*), Polish (*pl*), Russian (*ru*), Czech (*cs*), French (*fr*), German (*de*), Japanese (*ja*), and Chinese (*zh*). As target languages, we select English, Japanese, and Polish, as they also vary greatly in their morphology, grammar, and writing systems. The detailed rationale can be found in §B.

3.2 Translation with large language models

In this paper, we focus on translating the literary paragraphs in our dataset using large language models. More specifically, we use the GPT-3.5 `text-davinci-003` checkpoint, which has been further tuned to follow instructions based on human feedback (Ouyang et al., 2022). Hendy et al. (2023) demonstrate that GPT-3.5 produces translations of reasonable quality, though their focus was mainly at the sentence level. Since many LLMs, including GPT-3.5, are only accessible via black-box APIs, we adapt the model for translation via in-context learning (Brown et al., 2020).

Demonstration examples: We use few-shot prompting, in which a model is provided with a prompt consisting of five demonstrations. We manually curate the five demonstrations from literary texts for each of the 18 language pairs, resulting in 90 total demonstration examples. These demonstrations are sourced from novels that are *not* part of our translation dataset, resulting in potential differences in topic and style (see Table 9 in the §B for details). We further ensure that each set of five demonstrations includes both dialogues and narrative texts.

Prompting for translation: We consider the following three prompting strategies for GPT-3.5 that allow us to compare the model's abilities to translate with and without discourse-level context (see Table 1 for templates and §C for the exact prompts):

- **GPT-3.5 sentence-level translation without context (SENT):** Each sentence of the paragraph is translated in isolation of the others. To maintain consistency, we provide the same five *sentence*-level examples[10] in each prompt for the given source-target language pair.[11]
- **GPT-3.5 sentence-level translation *with* context (PARA_SENT):** Each sentence of the paragraph is translated in context. The model is provided with the entire source paragraph as input, where the sentence to be translated is wrapped in `<translate>` and `</translate>` tags, in addition to a partially-translated target paragraph. The demonstrations are also presented with the same tags. For each demonstration in the prompt, a sentence in a different position was chosen (e.g., from the beginning, middle, and end of the paragraph).
- **GPT-3.5 paragraph-level translation (PARA):** The entire source paragraph is passed into the model, and the output target paragraph is generated conditioned on this input (i.e., without any sentence tokenization). Demonstrations in the prompt are also *paragraphs*[12] of translations from the respective source language into the target language in question.[13]

[9] A paragraph with fewer sentences is not necessarily short: for example, in the German novel "An Inventory of Losses," sentences can be as long as 70 to 80 words, with the longest reaching 117 words. The distribution of sentences in paragraphs is provided in Figure 7 in §B.

[10] Sentence-level demonstrations for SENT are sampled from the demonstrations for paragraph-level translation.

[11] To ensure consistent quotation mark usage and enable a fair comparison with paragraph-level translations, quotation marks in sentence-level translations were manually adjusted.

[12] The examples for PARA and PARA_SENT configurations are necessarily lengthier. Due to the GPT-3.5 maximum context size, it is not always possible to include all five examples within the prompt. Consequently, around 10% of the data was translated using four or fewer examples.

[13] Initially, we experimented with GPT-3 by translating between two non-English languages using English as a pivot, as it is the primary language of the model. The model had access to the source text and its English translation. After manual evaluation and comparison to translations without a pivot language, we found no significant benefit in using English as the pivot. Consequently, we directly translated paragraphs into the target language. Refer to §H for details and results of this preliminary study.

SENT Demonstration Template
Original text in [SRC LANG]: source sentence Translation into [TRG LANG]: target sentence
PARA_SENT Demonstration Template
Original text in [SRC LANG]: source sentence 1 <translate> source sentence 2 </translate> source sentence 3 Translation into [TRG LANG]: target sentence 1 <translated> target sentence 2 </translated>
PARA Demonstration Template
Original text in [SRC LANG]: source paragraph Translation into [TRG LANG]: target paragraph

Table 1: Prompt templates for SENT, PARA_SENT, and PARA. The source text to translate and expected target outputs are underlined.

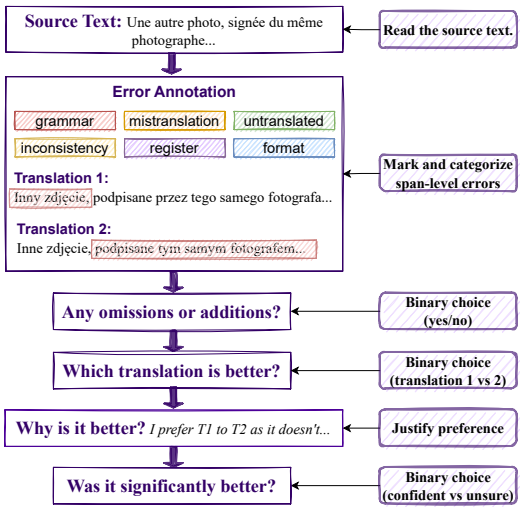


Figure 3: A description of the annotation process for a pair of candidate translations given a source paragraph. Note that our hired translators go through this pipeline for *three* different pairs per source paragraph, comparing PARA with SENT, PARA_SENT, and GTR.

Using Google Translate (GTR) as a baseline: In order to compare commercial-grade translation systems to LLM translators, we also translate all paragraphs in our dataset using Google Translate.[14] We opt for an off-the-shelf commercial system instead of a state-of-the-art system from, for instance, WMT competitions for two primary reasons. First, our experiments focus on *literary* translations. Given that WMT systems are predominantly evaluated on the news domain, it is uncertain which system would perform best, and some language pairs may not even be supported. Second, our main research question revolves around LLMs' ability to incorporate contextual information, rather than merely comparing their performance with state-of-the-art translation systems. We employ GTR as a reasonably robust baseline to assess the extent to which context can enhance MT quality, rather than asserting that LLMs outperform *all* traditional MT systems.

4 Evaluating document-level literary translation

How do we compare the translation quality of the systems described above? Automatic metrics such as BLEURT and COMET are untested on document-level inputs as well as literary texts, and as such we do not consider them reliable, although we do report them in §G.[15] Human evaluation is equally problematic, as direct assessments of translation quality (e.g., "rate the quality of this translation from 0-100") suffer from calibration issues that are exacerbated with longer texts (Karpinska et al., 2021). Thus, we opt for a human evaluation inspired by Multidimensional Quality Metrics (Lommel et al., 2014b, MQM), in which annotators mark and classify error spans within the translation. Specifically, for each of the 18 language pairs studied in this work, we hire translators to identify all span-level errors in two competing translations. For each evaluated pair, the annotators were also asked to choose the better translation and provide a free-form rationale. For each source paragraph, the translators make three binary judgments of which translation is higher quality: SENT vs PARA, PARA_SENT vs PARA, and GTR vs PARA.

Recruiting annotators: As our task is complex and requires fluency in both the source and target language, we hire *translators* to provide the annotations. We recruit 13 translators via the Upwork freelancing platform,[16] each of whom is a native speaker of English, Polish, or Japanese.[17] One

[14] All paragraphs were translated in January 2023 using the `GoogleTranslate` API. The system was provided entire paragraphs, which it likely partitioned and translated sentence-by-sentence.

[15] Automatic metrics developed specifically for document-level MT are also insufficient as they either work best with one-to-one sentence level alignments (Vernikos et al., 2022; Hendy et al., 2023) or are available only for English (Jiang et al., 2022).

[16] https://www.upwork.com/

[17] The annotators for Czech-Polish and Russian-English were both native speakers of the respective source languages

translator, hired directly, was a bilingual speaker of English and Polish with advanced knowledge of German; as such, she performed the *pl-en*, *de-en*, and *de-pl* evaluations. Evaluation of *ja-pl*, *pl-ja*, and *pl-en* texts was done by the first author in a collaboration with native speakers of Polish/Japanese to avoid any potential bias. Each translator was paid $2 per evaluated pair of candidate translations, with an additional $5 bonus to cover the time spent familiarizing themselves with the instructions. We asked them to compare three pairs of system translations (PARA vs. SENT, PARA vs. PARA_SENT, PARA vs. GTR) for 10 paragraphs per language pair[18]; as such, 180 total source paragraphs were used in our evaluations. Altogether, we paid approximately $12 per hour, with a total cost of $955.

Annotation task: First, we tasked the hired translators[19] with annotating a subset of MQM translation errors identified through a pilot analysis and annotation of the system's outputs. Specifically, we ask them to highlight spans within the candidate translations that contain errors belonging to any of the following error categories:

- **mistranslation:** [20] accuracy errors that occur when the wrong target word or phrase is chosen to represent content from the source text. In addition to canonical mistranslations, we also include *overly literal* translation errors that occur when systems nonsensically translate word-by-word into the target language.
- **grammar:** grammatical errors, such as errors in conjugation, declension, or wrong prepositions.
- **untranslated:** words or phrases that should have been translated into the target language but were either left in the source language or just transliterated into the target language.
- **inconsistency:** use of different terms to refer to the same entity, or different words where the same word should be used for stylistic reasons (e.g., "Kasia" and "Kate," "coat" and "jacket," or "bad" and "awful").
- **register:** a clear violation in the use of formal and informal language within the same text, only annotated in Japanese.[21]
- **format:** incorrect usage of punctuation (e.g., "." instead of "。").

After the span-level annotation is complete, we then ask the translators to further identify if any of the candidate translations contains significant content **additions** or **omissions** in relation to the source text.[22] Finally, they are asked to **choose the better translation** and provide a justification for their choice in two to five sentences. We instruct them to additionally mark whether their chosen translation is significantly superior, or if the decision was difficult because both translations are of roughly comparable quality (see Figure 3 and §D for details).

5 Results

In this section, we compare our different literary translation methodologies using both automatic metrics and aggregate statistics from the human evaluations. Overall, we observe that the PARA configuration outperforms competing methods across all evaluations and language pairs. These results demonstrate that GPT-3.5 effectively leverages paragraph-level context to produce better translations than sentence-level methods, and also that the less efficient sentence-by-sentence translation with context is (PARA_SENT) is unnecessary to achieve high translation quality.

5.1 Human evaluation also favors PARA

Figure 5 contains human preference results comparing PARA to SENT, PARA to PARA_SENT, and

and highly proficient in their respective target languages. They collaborated with native speakers of the target languages, who possessed a basic understanding of the source language, to complete their annotations.

[18]These paragraphs were randomly sampled from the 360 paragraphs. The entire set of 360 paragraphs was used for the automatic evaluation described in §G.

[19]They were presented with guidelines in their native language. The annotation task was performed using the Label-Studio annotation tool (Tkachenko et al., 2020-2022). See Figure 11 for the screenshot of the interface.

[20]We note that mistranslations in literary text are often not as grave as, for instance, in news articles. Human translators hold *poetic license*, which allows them to change some details to make the text more enjoyable for the reader. Is changing "bonito" into "tuna" incorrect? Or can it be perceived as a way to accommodate an English-speaking readership that is likely more familiar with the latter?

[21]We only annotate cases where the level of formality changes abruptly within the same paragraph. It is possible that a given character would be more likely to use formal language but an informal language is being employed. As long as this is consistent we do not consider it an error as this cannot be fully determined from the paragraph context.

[22]Note that this task was simplified to a binary choice – either there were serious omissions/additions or not. We did not ask the annotators to further annotate them due to the time restrictions.

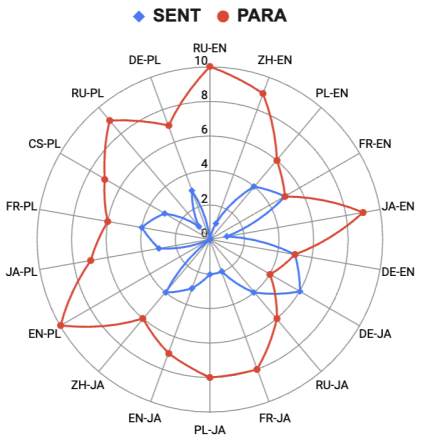


Figure 4: The distribution of translator preference judgments between sentence-level translation (SENT) and paragraph-level translation (PARA). PARA is preferred (i.e., more votes) in every language pair except *de-ja*, *fr-en* and *de-en*.

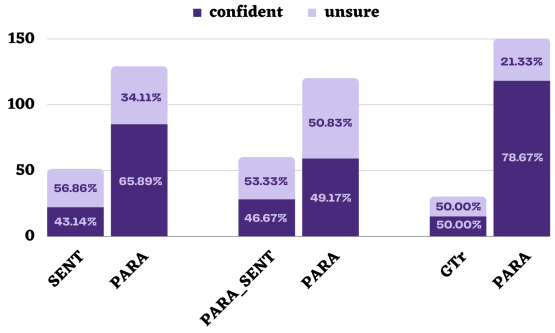


Figure 5: The number of votes for SENT vs PARA, PARA_SENT vs PARA, and GTR vs PARA along with rater confidence (*confident* or *unsure*). PARA is preferred to all competing methods. All differences are statistically significant at $p<.001$ (binomial test).

PARA to GTR, aggregated across all 18 language pairs studied in this paper (i.e., 180 votes per system comparison). Table 11 breaks down these results for each language pair, and we observe the same trends for the vast majority of pairs. Overall, the translators significantly favored PARA translations over the alternatives ($p<.001$, binomial test). Table 2 contains specific information about grammar and mistranslation errors split across the three target languages (see Table 16 and Table 17 for details), which we refer to in the discussion below.

PARA is clearly better than SENT: PARA is preferred by translators over SENT at a rate of 71.67% ($p<.001$, 95% CI [0.645, 0.781]). Additionally, when translators preferred PARA, they were usually confident in the decision (i.e., it was clearly better than SENT); even if we exclude all "unsure" votes, the preference for PARA translations remains significant at 79.44% ($p<.001$, 95% CI [0.705, 0.866]). The only language pair in which SENT is favored over PARA is *de-ja* (see Figure 4).[23] Overall, SENT produces 31% more mistranslations, 48.6% more grammar errors, 15 times more inconsistencies, and 3.5 times more register errors (Table 2).

PARA is clearly better than GTR: PARA translations are overwhelmingly preferred over those from Google Translate (GTR), with an 83.33% preference rate ($p<.001$, 95% CI [0.771, 0.885]). In the *fr-ja*, *pl-ja*, *zh-ja*, and *cs-pl* language pairs, PARA received *all* of the ten votes over GTR. Overall, GTR translations result in 58.18% more mistranslations, 35.24% more grammatical errors, over seven as many inconsistency errors, and ten times more register errors (see Table 2). §E contains more fine-grained comparisons of these two systems.

PARA is slightly preferred over PARA_SENT: Our evaluations show that PARA is better than PARA_SENT, but the gap is smaller than it is for the other two methods. PARA is still preferred at a 66.67% rate ($p<.001$, 95% CI [0.593, 0.735]). Both PARA and PARA_SENT produce a comparable number of mistranslations (*483* vs 462), grammar errors (*105* vs *113*), and inconsistencies (*2* vs *3*) (see Table 2). While PARA_SENT leaves around 22% more words untranslated, it appears to leverage the contexts and even occasionally selects better equivalents in the target language, as evidenced by translator comments. One major issue with PARA_SENT is that it occasionally repeats sentences, whereas PARA never does so.

6 Analyzing translation errors

The aggregate statistics from the previous section confirm that PARA-level translation via GPT-3.5 is the strongest literary translator of the methods that we study. Translations produced by PARA are favored by both automatic metrics and human translators, and it makes fewer errors than competing methods. In this section, we dive deeper into specific *types* of errors that are made within each high-level category (e.g., grammar, mistranslation), and we present examples of errors associated with

[23] This could be because the German novel *An Inventory of Losses* in our dataset contains the longest sentences of any book (45 tokens per sentence), and thus the intra-sentence context is likely more informative than in other books.

Type	Trg Lang	Para	Sent	Para_Sent	GTr
Mistranslation	En	88	109	82	155
	Ja	224	295	223	334
	Pl	171	229	157	275
	Total	*483*	*633*	*462*	*764*
Grammar	En	5	20	9	18
	Ja	43	49	38	65
	Pl	57	87	66	59
	Total	*105*	*156*	*113*	*142*
Inconsistency	En	0	5	0	1
	Ja	1	7	2	7
	Pl	1	19	1	7
	Total	*2*	*31*	*3*	*15*
Untranslated	En	13	5	14	6
	Ja	23	30	33	24
	Pl	23	16	25	4
	Total	*59*	*51*	*72*	*34*
Register	En	0	0	0	0
	Ja	7	25	13	71
	Pl	0	0	0	0
	Total	*7*	*25*	*13*	*71*
Format	En	0	*n/a*	*n/a*	1
	Ja	0	*n/a*	*n/a*	117
	Pl	0	*n/a*	*n/a*	8
	Total	*0*	*n/a*	*n/a*	*126*

Table 2: Total counts of all of the types of mistakes made by each of the four systems from our annotation. Overall, models with access to paragraph-level context commit fewer translation errors.

lack of context understanding made by Sent and GTr that are fixed by Para.

6.1 Language-specific grammatical errors

We analyze the types of grammatical errors that are made by the studied translation methods in all three target languages.[24] In summary, although GPT-3.5 is primarily trained on English, it is competitive with GTr at Polish and Japanese grammar proficiency. In fact, Para generates the fewest grammatical errors of any system, with a total of *97* for both languages, in contrast to *136* errors made by Sent, *101* errors by Para_Sent, and *122* errors by GTr (see Table 2). That said, *none* of these systems delivers translations devoid of grammatical inaccuracies, even for English.

English: Perhaps not surprisingly, translations into English contain fewer grammatical mistakes than Japanese or Polish (see Table 2). The most prominent mistakes in English are incorrect articles, which is most frequent with Sent and GTr. This is to be expected, as the choice between the definite and indefinite article in English depends heavily on the context. Other mistakes include wrong or omitted prepositions, wrong parts of speech, and incorrect word order (see Table 17).

[24]There are some differences in the paragraph lengths between the three target languages that should be taken into consideration when analyzing raw numbers. However, the general tendencies remain intact.

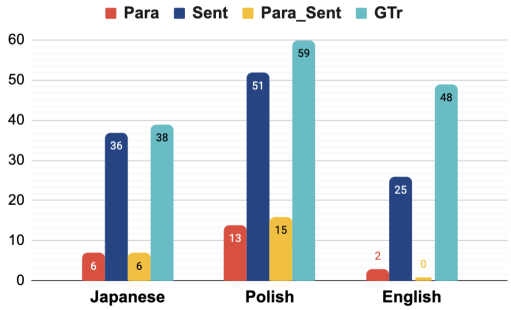


Figure 6: Quantification of mistranslations resulting from missing or misinterpreted paragraph-level context in Para, Sent, Para_Sent, and GTr systems, organized by the target language (Japanese, Polish, and English).

Japanese: Translations into Japanese contain considerably more mistakes. Most notably, the systems struggle with the correct choice of particle: Para and Sent produce twice as many mistakes in this regard than Para_Sent and GTr (see Table 17). Other mistakes include incorrect tense, verb finite form within the sentence, or incorrect word order, the latter of which is much more frequent in GTr than any of the GPT-3.5 translations.

Polish: GPT-3.5 exhibits more difficulty with Polish grammar than English or Japanese across all prompting strategies (see Table 2). It frequently generates incorrect gender, case, or prepositions (see Table 17). We also observe instances in which GPT-3.5 alters the gender of a noun, such as producing *grilla*, a non-existent feminine form, in place of the masculine *grill*, while accurately modifying all adjectives and verbs to match the novel feminine noun.[25]

6.2 Context-related errors

We manually classify *all* annotated mistranslations (2,324 instances) into subcategories, several of which include instances where the absence of discourse-level context is clearly a contributing factor (see Table 16 for detailed classification).[26] We also further analyze *all* translations in terms of content-related issues. Overall, we observe that context is indeed incorporated into the translations

[25]It is worth noting that *grilla* can also be also the genitive form of the masculine noun *grill*; however, the agreement of surrounding verbs and adjectives with the feminine noun suggests that the system likely treated the word as feminine.

[26]The initial classification was conducted on the first version of the dataset. After incorporating small corrections, we identified 18 more mistranslations that were not part of this analysis.

Type	Source	GPT-3.5 Sent Translation	GPT-3.5 Para Translation	Comment
Pronouns	Романы, как известно, печатались на разной бумаге [*paper*]. И гореть она [*she*] может по-разному. —Russian Source (from *Manaraga*)	Romany, jak wiadomo, drukowano na różnym papierze [*paper*]. I może ona [*she*] tęsknić na różne sposoby. —GPT-3.5 Sent (Polish)	Jak wiadomo, powieści drukowano na różnym papierze [*paper*]. I może on [*he*] palić się na różne sposoby. —GPT-3.5 Para (Polish)	"Paper" is a feminine noun in Russian and referred to as "she," whereas it is a masculine noun in Polish and should be referred to as "he," as in Para. The absence of context in Sent leads to an incorrect translation.
Cultural Nuances	「気が付かなくてすみません」「いやいや、(...). 古倉さんは毎日勤務なのに手を抜かないからねー！」 [lit. *Ms. Furukura works every day*] —Japanese Source (from *Convenience Store Woman*)	"I'm sorry I didn't notice." "No, no, (...). Furukura-san works hard every day without taking any shortcuts!" —GPT-3.5 Sent (English)	"I'm sorry I didn't notice." "No, no, (...). You work every day, but you never slack off!" —GPT-3.5 Para (English)	"Furukura-san" or "Miss Furukura" in the last source sentence is used instead of the second-person "you" as per Japanese convention. Translating this sentence without context into English results in a confusing translation (Sent) that implies that the speaker refers to some other "Furukura" rather than their listener. Para correctly translates "Furukura" as "you."
Ellipsis	„Ne, teď uděláš nádobí!" [*(you) will do the dishes!*] „Neudělám!" [*(I) won't do!*] „Uděláš!" [*(You) will do!*] —Czech Source (from *Crows*)	— Nie, teraz zrobisz zmywanie! [*(you) will do the washing*] — Nie zrobię! [*(I) won't do!*] — Zrobisz to! [*(You) will do it!*] —GPT-3.5 Sent (Polish)	— Nie, teraz umyjesz naczynia [*(You) will wash the dishes*]! — Nie umyję [*(I) won't wash*]! — Umyjesz [*(You) will wash*]! —GPT-3.5 Para (Polish)	Czech uses the same collocation as English, "*do* the dishes," which is invalid in Polish. Hence, the ellipses in the last two sentences in the source text require a broader context to be translated correctly. Para does it properly, translating both as "wash," while Sent unsurprisingly fails to choose the correct collocation.
Subject Ellipsis	When we were done, the lipstick went back into some mother's Fendi handbag. We watched her apply it, unaware. —English Source (from *A Children's Bible*)	Gdy skończyliśmy, szminka wróciła do jakiejś torebki Fendi należącej do matki. Patrzyliśmy, jak to robi, nieświadomi [*unaware (we)*] tego. —GPT-3.5 Sent (Polish)	Kiedy skończyliśmy, szminka wróciła do torebki Fendi jakiejś matki. Patrzyliśmy, jak ją nakłada, nieświadoma [*unaware (she)*] naszych działań. —GPT-3.5 Para (Polish)	Only from the broader context we can deduce that "unaware" refers to the mother, not the "we" (referring to children) watching her. Para correctly attributes the state of being "unaware" to the mother, which is exhibited by its usage of the singular feminine form of the adjective. In contrast, Sent mistranslates it using the plural masculine form of the adjective "unaware," which implies that it refers to "we" rather than the "mother."
Consistency	Alles zu vergessen, ist gewiss schlimm [*bad*]. Noch schlimmer [*worse*] ist, nichts zu vergessen (...). —German Source (from *An Inventory of Losses*)	すべてを忘れることは確かに悲惨な[*tragic*]ことです。さらに悪い[*worse*]のは、何も忘れないことです。 —GPT-3.5 Sent (Japanese)	すべてを忘れることは確かに悪い[*bad*]ことです。もっと悪い[*worse*]ことは、何も忘れないことです。 —GPT-3.5 Para (Japanese)	The German source translates into English as "To forget everything is *bad*, certainly. *Worse* still is to forget nothing." It is arguably important for the translation to repeat the same word which is an equivalent of the German "schlimm" ("bad"). Para does it well, translating both as 悪い, or "bad," while Sent uses two different words, "tragic" and "bad" which results in inconsistent translation.
Polysemy	Все прошло хорошо. Книга прочитана идеально – не быстро и не медленно, минимум дыма. Классика. Я был в форме [*in shape*]. —Russian Source (from *Maranaga*)	Wszystko poszło dobrze. Książka została przeczytana idealnie – nie szybko i nie wolno, minimalna ilość dymu. Klasyka. Byłem w mundurze [*in uniform*]. —GPT-3.5 Sent (Polish)	Wszystko poszło dobrze. Książka przeczytana idealnie – nie szybko i nie wolno, minimalna ilość dymu. Klasyka. Byłem w formie [*in shape*]. —GPT-3.5 Para (Polish)	The ambiguity stems here from multiple meanings of the Russian noun форма , which can mean either "shape" or "uniform." Since one can be "in shape" as well as "in a uniform", only from the context it becomes clear which meaning was intended by the author. Para translates it correctly as "shape" while Sent mistranslates it as "uniform."
Appropriateness	「あー、あと煙草の５番を一つ」「かしこまりました」 [lit. *(I) understood*] —Japanese Source (from *Convenience Store Woman*)	"Ah, and one pack of cigarettes, number five." "Understood." —GPT-3.5 Sent (English)	"Ah, and one pack of cigarettes, number five." "Right away." —GPT-3.5 Para (English)	This conversation is between a clerk and a customer. The Japanese expression かしこまりました is an honorific that literally means "understood." However, when choosing the best equivalent, the translator needs to consider the situation at hand to best reflect its meaning in the target language. "Understood" in Sent is technically correct, but it is an unfortunate word choice for the clerk to employ. On the other hand, "right away" in Para fits much better in the context of this conversation.

Table 3: Examples of different context-related issues observed in Sent translations, which are fixed in the corresponding Para translations. Phrases that exemplify these issues are highlighted in purple, and English glosses are provided in [*square brackets*].

for both Para and Para_Sent outputs, which results in fewer context-dependent issues (see Figure 6). More specifically, we observe that Para produces translations that leverage the context resulting in mostly correct translations of pronouns, ellipsis, cultural nuances, and polysemous words and phrases; Table 3 contains specific examples and discussion of each. Para is also more consistent and appropriate in vocabulary usage than Sent. All cases are further analyzed in §E.2.

7 Conclusion

In this paper, we demonstrate that LLMs leverage paragraph-level context to produce translations that are more coherent and enjoyable than sentence-by-sentence translation while containing fewer mistranslations and grammatical issues. Our evaluations reveal that professional translators prefer paragraph-level translations over both sentence-level translations produced by the same language model, and also to those generated by an off-the-shelf commercial system (GTr). We release our dataset and error annotations to help facilitate the development of new evaluation methodologies and automatic metrics for document-level machine translation. Finally, a full-length novel extends far beyond the confines of paragraph-level translation. In future work, we will focus on integrating individual paragraphs into cohesive chapters, which can then be expanded to encompass the entire novel.

8 Limitations

So far, we have shown that GPT-3.5 leverages paragraph-level context to produce translations that

are better than those produced by sentence-level counterparts (SENT vs PARA). However, there are still many issues with PARA's translations. From the annotations and translators' comments, we observe that PARA suffers from occasional omissions of content from the source paragraph to a greater extent than SENT and GTR (see §D). Moreover, PARA still makes a sizeable number of mistranslations and grammatical errors, though fewer than SENT or GTR. These issues seem to be only partially mitigated by employing GPT-4 (see §F). Finally, it is important to acknowledge that the languages covered in the current study are either mid or high-resource. Performance might be much worse when translating from or into a low-resource language such as Zulu or Armenian.[27]

Ethical considerations

Translating with LLMs: The rise of large language models has also brought many ethical concerns to the forefront of NLP research (Blodgett et al., 2020; Bender et al., 2021). LLMs encode biases and exhibit toxicity, and these behaviors can be exacerbated by unconstrained prompting (Gehman et al., 2020; Costa-jussà et al., 2022). Further ethical concerns arise in the context of machine translation, particularly *literary* translation, where multiple stakeholders – the author, the translator, and the audience – are involved (Taivalkoski-Shilov, 2019a). Low-quality output can influence the perception of the author's work, impair the reader's linguistic abilities, and hinder the transfer of ideas to the target language, while overrelying on machine translation can possibly threaten the role of human translators (Drugan, 2013; Ning and Domínguez, 2016; Taivalkoski-Shilov, 2019a). On the other hand, machine translation employed *responsibly* as an auxiliary tool holds the potential to alleviate the translator's cognitive burden (O'Brien, 2012) and make the author's work accessible to a broader audience more swiftly (Besacier, 2014). Contrary to the predictions in Eloundou et al. (2023), we do not view large language models as a substitute for human translators, but rather as a means to assist translators in their work.

Human Evaluation: The experiments involving human translators were reviewed by the IRB, and all involved translators gave their written consent to disclose their annotations, comments, and preference choices. In recognizing contributions, our acknowledgments only include the names of those translators who explicitly gave their consent to be acknowledged by their full name in this publication.

Data Copyrights: We use and make public only about 2% of the text from each of the original novels. This number was determined after consulting domain experts at the HathiTrust (https://www.hathitrust.org/) and qualifies as fair use (up to 10% of a text can generally be considered fair use).

[27] For instance, our initial experiments with translations into low-resource languages show that GPT-3.5 (ChatGPT) suffers from repetition when translating into Hausa.

Acknowledgements

First and foremost, we would like to express our gratitude to the translators hired mostly on Upwork: Malgorzata Szymczak (*fr-pl*), Kinga Przekota (*ru-pl*), Michal Sikora (*cs-pl*), Paula Kurzawska (*de-pl*, *de-en*, *pl-en*), Kristy Darling Finder (*fr-en*), Timothy Shostak (*ja-en*), Shun Enoki (*zh-ja*), Takanori Kurokawa (*fr-ja*), Yoshiko Kikawa (*en-ja*), Shinnosuke Kasahara (*ru-ja*), and all those who wish to remain anonymous. We encourage any machine translation researchers working on these language pairs to contact these translators for human evaluations.

We would also like to show our appreciation to Jan Wislicki, Tom Gally, Nader Akoury, Kalpesh Krishna, Simeng Sun, Katherine Thai, and the entire UMass NLP group for insightful discussion, which helped to shape this project.

Furthermore, we would like to express our gratitude to the reviewers for their constructive feedback and valuable suggestions.

Finally, we would like to thank Sergiusz Rzepkowski (*pl*), Paula Kurzawska (*pl, en*), Hiroshi Iida (*ja*), Grégory Fleurot (*fr*), Peyton Bowman (*en*), Simeng Sun (*zh*), Igor Zapala (*pl, de*), Marvin Hoffmann (*de*), Kinga Przekota (*pl, ru*), and Yuki Mori (*ja*) for further consultations on their respective native languages.

This project was partially supported by awards IIS-1955567 and IIS-2046248 from the National Science Foundation (NSF) as well as an award from Open Philanthropy.

References

Ruchit Agrawal, Marco Turchi, and Matteo Negri. 2018. Contextual Handling in Neural machine Translation:

Look Behind, Ahead and on Both Sides. In *21st Annual Conference of the European Association for Machine Translation*, pages 11–20.

Rachith Aiyappa, Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2023. Can we trust the evaluation on ChatGPT?

R.H. Baayen, D.J. Davidson, and D.M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Laurent Besacier. 2014. Machine translation for litterature: a pilot study (traduction automatisée d'une oeuvre littéraire: une étude pilote) [in French]. In *Proceedings of TALN 2014 (Volume 2: Short Papers)*, pages 389–394, Marseille, France. Association pour le Traitement Automatique des Langues.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Bilingual Book Maker. 2023. Make bilingual epub books Using AI translate (GitHub). `https://github.com/yihong0618/bilingual_book_maker`. [Accessed 05-Apr-2023].

Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for needles in a haystack: On the role of incidental bilingualism in palm's translation capability.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Marine Carpuat and Michel Simard. 2012. The Trouble with SMT Consistency. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 442–449, Montréal, Canada. Association for Computational Linguistics.

Sheila Castilho. 2021. Towards Document-Level human MT Evaluation: On the Issues of Annotator Agreement, Effort and Misevaluation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 34–45, Online. Association for Computational Linguistics.

Kent K. Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, memory: An archaeology of books known to chatgpt/gpt-4.

Junxuan Chen, Xiang Li, Jiarui Zhang, Chulun Zhou, Jianwei Cui, Bin Wang, and Jinsong Su. 2020. Modeling Discourse Structure for Document-level Neural Machine Translation. In *Proceedings of the First Workshop on Automatic Simultaneous Translation*, pages 30–36, Seattle, Washington. Association for Computational Linguistics.

Andrew Chesterman. 1997. *Memes of Translation*. Benjamins Translation Library. Benjamins (John) North America, Amsterdam, Netherlands.

Trevor Cohn and Mirella Lapata. 2007. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 728–735, Prague, Czech Republic. Association for Computational Linguistics.

Marta R. Costa-jussà, Eric Smith, Christophe Ropers, Daniel Licht, Javier Ferrando, and Carlos Escolano. 2022. Toxicity in Multilingual Machine Translation at Scale.

Chenchen Ding, Masao Utiyama, and Eiichiro Sumita. 2014. Document-level re-ranking with soft lexical and semantic features for statistical machine translation. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: MT Researchers Track*, pages 110–123, Vancouver, Canada. Association for Machine Translation in the Americas.

Joanna Drugan. 2013. *Quality in professional translation*. Bloomsbury Advances in Translation. Continuum Publishing Corporation, New York, NY.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models.

Yukun Feng, Feng Li, Ziang Song, Boyuan Zheng, and Philipp Koehn. 2022. Learn to Remember: Transformer with Recurrent Memory for Document-level Machine Translation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1409–1420, Seattle, United States. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2022. Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based Phrase-level Prompting of Large Language Models for Machine Translation.

Nuno M. Guerreiro, Duarte Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in Large Multilingual Translation Models.

Chao Han. 2020. Translation quality assessment: a critical methodological review. *The Translator*, 26(3):257–273.

Christian Hardmeier. 2012. Discourse in Statistical Machine Translation. *Discours*, (11).

Christian Hardmeier, Sara Stymne, Jörg Tiedemann, and Joakim Nivre. 2013. Docent: A Document-level Decoder for Phrase-Based Statistical Machine Translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 193–198, Sofia, Bulgaria. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation.

Klára Jágrová and Tania Avgustinova. 2023. Intelligibility of highly predictable polish target words in sentences presented to czech readers. In *Computational Linguistics and Intelligent Text Processing*, pages 110–125. Springer Nature Switzerland.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine translation benefit from larger context?

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An Automatic Evaluation Metric for Document-level Machine Translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States. Association for Computational Linguistics.

Yuchen Eleanor Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Mrinmaya Sachan, and Ryan Cotterell. 2023. Discourse-centric evaluation of document-level machine translation with a new densely annotated parallel corpus of novels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7853–7872, Toronto, Canada. Association for Computational Linguistics.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is ChatGPT A Good Translator? Yes with GPT-4 As The Engine.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233, Florence, Italy. Association for Computational Linguistics.

Xiaomian Kang, Yang Zhao, Jiajun Zhang, and Chengqing Zong. 2020. Dynamic context selection for document-level neural machine translation via reinforcement learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2242–2254, Online. Association for Computational Linguistics.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The Perils of Using Mechanical Turk to Evaluate Open-ended Text Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. DEMETR: Diagnosing Evaluation Metrics for Translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9540–9561, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large Language Models Are State-of-the-Art Evaluators of Translation Quality.

Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13):1–26.

Russell V. Lenth. 2023. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.8.5.

Arle Lommel, Maja Popovic, and Aljoscha Burchardt. 2014a. Assessing inter-annotator agreement for translation error annotation. Reykjavik, Iceland. Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 14).

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014b. Multidimensional Quality Metrics (MQM) : A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica*, pages 0455–463.

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A Systematic Comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

Hongyuan Lu, Haoyang Huang, Dongdong Zhang, Haoran Yang, Wai Lam, and Furu Wei. 2023. Chain-of-dictionary prompting elicits translation in large language models.

Elman Mansimov, Gábor Melis, and Lei Yu. 2021. Capturing document context inside sentence-level neural machine translation models with self-training. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 143–153, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Lucía Molina and Amparo Hurtado Albir. 2004. Translation techniques revisited: A dynamic and functionalist approach. *Meta*, 47(4):498–512.

Wang Ning and César Domínguez. 2016. Comparative literature and translation: A cross-cultural and interdisciplinary perspective. In Yves Gambier and Luc van Doorslaer, editors, *Border crossings. Translation studies and other disciplines*, pages 287–308. John Benjamins.

Sharon O'Brien. 2012. Translation as human–computer interaction. *Translation Spaces*, 1:101–122.

OpenAI. 2022. Introducing ChatGPT.

OpenAI. 2023. GPT-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Dorota Pawlak. 2023. ChatGPT for Translators: How to Use the Tool to Work More Efficiently?

Matt Post and Marcin Junczys-Dowmunt. 2023. Escaping the sentence-level paradigm in machine translation.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are References Really Needed? Unbabel-IST 2021 Submission for the Metrics Shared Task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

Juan C. Sager. 1998. What Distinguishes Major Types of Translation? *The Translator*, 4(1):69–89.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Kristiina Taivalkoski-Shilov. 2019a. Ethical issues regarding machine(-assisted) translation of literary texts. *Perspectives*, 27(5):689–703.

Kristiina Taivalkoski-Shilov. 2019b. Free indirect discourse: an insurmountable challenge for literary MT systems? In *Proceedings of the Qualities of Literary Machine Translation*, pages 35–39.

Xin Tan, Longyin Zhang, Deyi Xiong, and Guodong Zhou. 2019. Hierarchical modeling of global context for document-level neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1576–1585, Hong Kong, China. Association for Computational Linguistics.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Craig Thomson, Ehud Reiter, and Barkavi Sundararajan. 2023. Evaluating factual accuracy in complex data-to-text. *Computer Speech & Language*, 80:101482.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Antonio Toral and Andy Way. 2015. Machine-assisted translation of literary text. *Translation Spaces*, 4(2):240–267.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 484–491, Rochester, New York. Association for Computational Linguistics.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly Easy Document-level MT Metrics: How to Convert Any Pretrained Metric Into a Document-Level Metric. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2022. Prompting palm for translation: Assessing strategies and performance.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models.

Biao Zhang, Ankur Bapna, Melvin Johnson, Ali Dabirmoghaddam, Naveen Arivazhagan, and Orhan Firat. 2022. Multilingual document-level translation enables zero-shot transfer from sentences to documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4176–4192, Dublin, Ireland. Association for Computational Linguistics.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT.

Zaixiang Zheng, Xiang Yue, Shujian Huang, Jiajun Chen, and Alexandra Birch. 2020. Towards making the most of context in neural machine translation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3983–3989. International Joint Conferences on Artificial Intelligence Organization. Main track.

Appendix

A Background

In this section of the appendix, we survey the existing approaches to document-level machine translation, which do not involve prompting LLMs.

Existing approaches to document-level translation: Before the rise of neural machine translation, several attempts were made to incorporate discourse-level phenomena into statistical machine translation systems (Hardmeier, 2012; Carpuat and Simard, 2012; Hardmeier et al., 2013; Ding et al., 2014). Neural MT systems condition sentence-by-sentence translation on discourse-level context via concatenation models (Tiedemann and Scherrer, 2017; Jean et al., 2017; Agrawal et al., 2018; Junczys-Dowmunt, 2019; Lopes et al., 2020), hierarchical models (Miculicich et al., 2018; Tan et al., 2019; Chen et al., 2020; Zheng et al., 2020), multi-pass models (Mansimov et al., 2021), dynamic context models (Kang et al., 2020), multi-source models (Zhang et al., 2018; Feng et al., 2022), and transfer learning approaches (Zhang et al., 2022). Despite sometimes obtaining clear gains from discourse-level context (Voita et al., 2019), the machine translation community has not made much progress on this problem, particularly for non-English language pairs, due largely to the scarcity of parallel document-level corpora (Zhang et al., 2022). This problem has been partially addressed by introducing a pivot language (Cohn and Lapata, 2007; Utiyama and Isahara, 2007), but this approach can also lead to substantial information loss.

B Dataset

In this section of the appendix, we first discuss the rationale for the source and target language selection. Then we provide more details on the selection of the paragraphs. Finally, we provide details about the number of tokens and sentences in the source text and different translations.

Target language selection: We select English, Japanese, and Polish as the target languages of our study, as these languages differ considerably in many linguistic aspects. English is an analytic language that is widely spoken and extensively studied in the field of natural language processing, and it serves as the primary pretraining language of most large language models, including GPT-3.5.[28] In contrast, both Japanese and Polish are comparatively under-explored. Japanese is an agglutinative language that employs three distinct writing systems: Kanji, Hiragana, and Katakana. As a high-context language, the translation of Japanese texts necessitates a profound comprehension of context and cultural nuances, rendering it a compelling choice for testing the limits of LLMs' translation capabilities. Polish, on the other hand, is a fusional language characterized by a rich morphological system. Its complex word forms, grammatical gender, conjugation, and declension make it an apt choice for testing the accuracy and robustness of LLMs.[29]

Source language selection: As source languages, we select English (*es*), Polish (*pl*), Russian (*ru*), Czech (*cs*), French (*fr*), German (*de*), Japanese (*ja*), and Chinese (*zh*). These languages belong to a diverse array of language families – Indo-European (Romance, Germanic, Slavic), Sino-Tibetan, and Japonic – each with distinctive morphological traits – fusional, agglutinative, and analytic. Moreover, they employ a variety of writing systems such as the Latin alphabet, the Cyrillic alphabet, Hanzi, and Kanji/Hiragana/Katakana (see Table 4 for details). Finally, we carefully select source-target language pairs to ensure that our study encompasses both linguistically similar and dissimilar languages. For example, we paired *cs-pl*, as these languages are characterized by only 10% lexical distance[30] and have similar syntactic structures (Jágrová and Avgustinova, 2023). Conversely, we also include *ja-pl*, as the two languages have very little lexical overlap, vastly different grammars, and utilize distinct writing systems.

Choosing paragraphs: The selection of a particular paragraph was semi-random, with certain considerations in mind during the sampling process. We prioritized the following criteria: (1) for each

[28] As of 2020, the reported distribution of languages featured in the present study within the GPT-3 training data was as follows: English – 92.647% (1st), French – 1.818% (2nd), German – 1.469% (3rd), Russian – 0.188% (9th), Polish – 0.155% (11th), Japanese – 0.111% (15th), Chinese – 0.099% (17th), Czech – 0.071% (18th) (see https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_word_count.csv). The current GPT-3.5 `text-davinci-003` model is reported to incorporate data up to June 2021 and it is unclear what texts or languages were added to the original training data https://platform.openai.com/docs/models/gpt-3-5.

[29] The first author is fluent in all three target languages.

[30] i.e., the percentage of non-cognates in the language pair.

source language we sample paragraphs so that there is a combination of dialogue and narrative texts; (2) the paragraph should be reasonably intelligible to a human translator without additional context; and (3) alignment between the source paragraph and human translation should be feasible, meaning no major content rearrangement *across* paragraphs.

Nonetheless, meeting all these requirements was not always possible. For instance, the source text of *Convenience Store Woman* (*ja*) is mostly written in the first-person narrative. Since Japanese does not encode the speaker's gender in the verb forms, it is often impossible to determine whether the narrator is a male or a female. In cases where it was impossible to determine the gender of the character we instructed translators to accept *either* option, provided that the translation remained consistent within the given paragraph (i.e., the gender did not change within the paragraph).

Dataset statistics: The dataset used for this study contains 360 source paragraphs with their corresponding human translations.[31] We further report the following statistics: (1) the number of sentences in the source text, as per manual sentence tokenization, along with the number of tokens in the sources and text and each translation (Table 5), (2) the number of sentences in the source text, human translation, and each machine translation as tokenized with SpaCy (Table 6), and (3) the distribution of sentences in paragraphs (Figure 7).

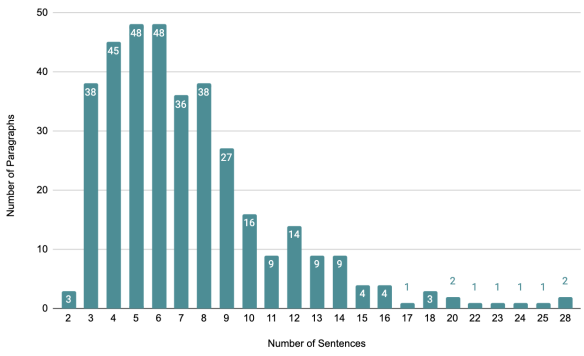


Figure 7: Distribution of sentences in the sampled paragraphs. The paragraphs were sentencized manually.

C Prompting for Translation

D Human Evaluation

In this section, we provide some further details about the human evaluation with a focus on the error annotation. First, discuss the issue of subjectivity in error annotation. Next, we explain some choices we had to make when annotating "inconsistency" and "format" errors. Finally, we present some details about the translators hired for the evaluation task.

Error annotation: Annotating and classifying errors in translations is inherently subjective (Lommel et al., 2014a; Han, 2020). For instance, translating French "corsage" ("bodice") as a "blouse" can be seen as either a mistranslation or a permissible deviation from the original text; this is, in fact, how the "corsage" was translated by the human translator in our data.

Furthermore, sometimes there are multiple ways of annotating errors (Thomson et al., 2023). Consider the following example:

> (1) We had to hide the running, though, in case our haste betrayed us, so truer to say we slipped out quietly. When one of my parents appeared, my technique was: pretend to catch sight of someone in the next room. **Move in a natural manner toward this figment of my imagination, making a purposeful face.**
>
> —English Source (from *A Children's Bible*)

The translation of the last sentence in (1) into Polish as an imperative can be considered a mistranslation. We would hypothesis that the system misinterpreted the source as an imperative form. However, using the infinitive form of the verb in the translation is less clear and raises questions about whether it is a mistranslation or a grammatical error. The distinction between the two lies in the point at which the mistake was made. If the original sentence was understood correctly but the resulting translation was ungrammatical, then it is a grammatical error. On the other hand, if the use of the infinitive form resulted from interpreting "move" as an infinitive, it may be considered a mistranslation as well.

Inconsistency: For marking the "inconsistency" errors we decided to the take *minimal* approach. For instance, is the same person is referred to in the translation as both "Piotr" and "Peter" we would

[31] See Table 7 for the list of novels included in the dataset and Table 8 for examples of aligned paragraphs.

LANGUAGE	LANGUAGE FAMILY	MORPHOLOGICAL FEATURES	WRITING SYSTEM
ENGLISH	Indo-European (Germanic)	Analytic	Latin Alphabet
GERMAN	Indo-European (Germanic)	Fusional	Latin Alphabet
FRENCH	Indo-European (Romance)	Fusional	Latin Alphabet
POLISH	Indo-European (Slavic)	Fusional	Latin Alphabet
CZECH	Indo-European (Slavic)	Fusional	Latin Alphabet
RUSSIAN	Indo-European (Slavic)	Fusional	Cyrillic
JAPANESE	Japonic	Agglutinative	Kanji / Hiragana / Katakana
CHINESE	Sino-Tibetan	Analytic	Hanzi

Table 4: Details on languages included for the current study.

LANG	#SENT	SRC	HUM	PARA	SENT	PARA_SENT	GTR
cs-pl	163	2,154	2,027	2,122	2,123	2,259	2,065
de-pl	153	3,172	2,997	2,785	2,899	2,835	2,764
ru-pl	170	2,350	2,471	2,467	2,463	2,458	2,375
ja-pl	111	2,627	1,855	1,782	1,907	1,830	1,800
en-pl	127	1,702	1,526	1,444	1,513	1,483	1,462
fr-pl	119	3,253	2,789	2,641	2,673	2,654	2,543
de-ja	75	3,530	5,329	4,807	5,116	4,652	4,703
en-ja	176	1,959	2,617	2,538	2,653	2,617	2,634
zh-ja	194	2,998	4,124	3,861	4,249	3,957	3,978
ru-ja	193	2,539	4,753	3,982	4,348	4,088	3,921
fr-ja	195	2,510	3,426	3,110	3,355	3,106	2,958
pl-ja	188	1,953	2,944	3,083	3,418	3,199	2,972
ja-en	111	2,622	2,293	2,062	2,322	2,257	2,140
pl-en	148	2,696	3,430	3,234	3,290	3,273	3,213
ru-en	117	1,693	2,008	2,029	2,056	2,028	2,019
fr-en	120	3,253	3,123	3,067	3,150	3,064	3,098
de-en	153	3,172	3,346	3,361	3,413	3,325	3,314
zh-en	127	2,235	2,002	2,427	2,396	2,351	2,360
Total	**2,640**	**46,418**	**53,060**	**50,802**	**53,344**	**51,436**	**50,319**

Table 5: Number of sentences in the source text sentencized manually (#SENT) along with the number of tokens in the human reference (HUM) and different machine translations (PARA, SENT, PARA_SENT, GTR). All translations were tokenized using SPACY[32] with the large model for each of the three target languages (Polish, Japanese, and English). All source texts were tokenized with STANZA (Qi et al., 2020) as SPACY does not include models for all target languages.

Original text in Japanese:
「そういうのは実際には起こらないの？」

Translation into Polish:
– To się w rzeczywistości nie zdarza?

(...)

Original text in Japanese:
「いらっしゃいませ！」

Translation into Polish:

Figure 8: An example of prompt for SENT translations with one demonstration and a text to translate.

LANG	SOURCE	TARGET	PARA	SENT	PARA_SENT	GTR
cs-pl	168	177	167	169	181	168
de-en	155	182	166	167	164	155
de-ja	69	133	135	121	117	132
de-pl	155	170	166	167	169	157
en-ja	169	168	166	161	169	169
en-pl	131	127	130	132	130	131
fr-en	122	138	126	122	124	123
fr-ja	193	199	207	220	185	201
fr-pl	122	125	125	125	126	123
ja-en	101	120	116	116	116	111
ja-pl	101	127	117	115	118	108
pl-en	148	156	149	145	151	145
pl-ja	189	153	174	196	178	191
ru-en	123	119	121	124	121	123
ru-ja	144	155	158	161	164	196
ru-pl	168	172	170	171	172	172
zh-en	127	130	146	141	140	135
zh-ja	195	234	225	229	215	202
TOTAL	**2,580**	**2,785**	**2,764**	**2,782**	**2,740**	**2,742**

Table 6: Number of sentences in the source text and each translation. The data was sentencized with SPACY. As evident from the data and manual inspection of translations the translations may result in a very different number of sentences as a result of splits and merges. We observe that about 55% of the data potentially lacks one-to-one correspondence.

Original text in Czech:
V hospodě U kalicha seděl jen jeden host. Byl to civilní strážník Bretschneider, stojící ve službách státní policie. **<translate>**Hostinský Palivec myl tácky a Bretschneider se marně snažil navázat s ním vážný rozhovor.**</translate>**

Translation into Polish:
W gospodzie „Pod Kielichem" siedział tylko jeden gość. Był to wywiadowca Bretschneider, będący na służbie policji państwowej. **<translated>**Gospodarz Palivec zmywał podstawki, a Bretschneider daremnie usiłował wyciągnąć go na poważną rozmowę.**</translated>**

(...)

Original text in Czech:
„Je to fakt dobrý," řekl mi Frodo, když se na můj výkres koukal. **<translate>**A skoro to vypadalo, že je z toho obrázku překvapenej.**</translate>** Pak se ptal, jestli maluju i doma, tak jsem odpověděla, že jo. Nejradši bych mu řekla i to, že chci být malířkou, ale mamka pořád opakuje, že je to blbost, že člověk musí dělat něco pořádnýho, jako třeba ona dělá knihovnici v knihovně, tak jsem raději mlčela, aby si nemyslel, že mám blbý nápady. Chvíli na výkres ještě koukal a otáčel ho ze všech stran a pak šel dál a koukal na obrázek Lindy, která nakreslila takový docela pěkný jabko.

Translation into Polish:
– To jest naprawdę dobre – powiedział Frodo, patrząc na moją pracę. **<translated>**

Figure 9: An example of prompt for PARA_SENT translations with one demonstration and a text to translate.

mark only the one that is less frequent. If "Piotr" appears *once* in the paragraph, while "Peter" is used *twice*, "Piotr" would be annotated as being inconsistent. The same strategy was applied for "register" errors, such as when both polite and casual forms were acceptable, but the translation used them randomly.

Format: We did not label "format" errors for the SENT and PARA_SENT translations, as we manually corrected the quotation marks during post-processing of the translations. This manual correction was done to ensure that SENT and PARA_SENT could be compared to PARA without relying too heavily on simple heuristic (i.e., incorrect usage of the quotation marks).

Translators: The translators in this study were hired on a freelancing platform, Upwork. We interviewed all translators prior to the task to assure

Book title	Author	Translator(s)	Language		Year Published	
			Source	Target	Translation	Original
A Children's Bible	Lydia Millet	Aga Zano	*en*	*pl*	2022	2020
What Can You See From Here	Mariana Leky	Agnieszka Walczy	*de*	*pl*	2021	2017
The Years	Annie Ernaux	Krzysztof Jarosz & Magdalena Budzińska	*fr*	*pl*	2022	2008
Manaraga	Wladimir Sorokin	Agnieszka Lubomira Piotrowska	*ru*	*pl*	2018	2017
Crows	Petra Dvorakova	Mirosław Śmigielski	*cs*	*pl*	2020	2020
Convenience Store Woman	Sayaka Murata	Dariusz Latoś	*ja*	*pl*	2019	2016
Sixteen Horses	Greg Buchanan	Fuji Yoshiko	*en*	*ja*	2022	2021
An Inventory of Losses	Judith Schalansky	Naoko Hosoi	*de*	*ja*	2022	2018
Dear Reader	Paul Fournel	Kei Takahashi	*fr*	*ja*	2022	2011
The Shooting Party	Anton Chekhov	Takuya Hara	*ru*	*ja*	2022	1884
Sword of Destiny	Andrzej Sapkowski	Yasuko Kawano	*pl*	*ja*	2022	1992
Bare burial	Fang Fang	Shin'ichi Watanabe	*zh*	*ja*	2022	2016
What Can You See From Here	Mariana Leky	Tess Lewis	*de*	*en*	2021	2017
The Years	Annie Ernaux	Alison L. Strayer	*fr*	*en*	2017	2008
The Story of a Life	Konstantin Paustovsky	Douglas Smith	*ru*	*en*	2022	1956
The Books of Jacob	Olga Yokarczuk	Jennifer Croft	*pl*	*en*	2022	2014
Convenience Store Woman	Sayaka Murata	Ginny Tapley Takemori	*ja*	*en*	2018	2016
Cocoon	Zhang Yueran	Jeremy Tiang	*zh*	*en*	2022	2018

Table 7: Details of the translated novels used in our study. In cases where the same novel is used for multiple target languages (e.g., "The Years"), identical source paragraphs are extracted to enable comparisons across language pairs. These novels exhibit distinct differences beyond just their source languages. For instance, "What Can You See From Here" presents a philosophical exploration of life and death, while "Sword of Destiny" is a fantasy story part of "The Witcher" saga.

Original text in Japanese:
直子は立ちどまった。僕も立ちどまった。彼女は両手を僕の肩にあてて正面から、僕の目をじっとのぞきこんだ。彼女の瞳の奥の方ではまっ黒な重い液体が不思議な図形の渦を描いていた。そんな一対の美しい瞳が長いあいだ僕の中をのぞきこんでいた。それから彼女は背のびをして僕の頬にそっと頬をつけた。それは一瞬胸がつまってしまうくらいあたたかくて素敵な仕草だった。

Translation into Polish:
Naoko zatrzymała się. Ja też stanąłem. Położyła mi ręce na ramionach i zajrzała uważnie w oczy. W jej ciemnych jak atrament źrenicach tworzyły się przedziwne wirujące wzory. Te piękne oczy długo badały moje serce. Potem wyprostowała się i przytuliła policzek do mojego. To był cudowny ciepły gest, aż mi serce na chwilę zamarło.

(...)

Original text in Japanese:
店長は30歳の男性で、常にきびきびとしている。口は悪いが働き者の、この店で８人目の店長だ。２人目の店長はサボり癖があり、４人目の店長は真面目で掃除好きで、６人目の店長は癖のある人で嫌われ、夕勤が全員一気に辞めるというトラブルになった。８人目の店長は比較的アルバイトからも好かれ、自分が体を動かして働くタイプなので、見ていて気持ちがいい。７人目の店長は気弱すぎて夜勤になかなか注意ができずに店がぼろぼろになってしまったので、少し口が悪くてもこれくらいのほうが働きやすいと、８人目の店長を見ると思う。

Translation into Polish:

Figure 10: An example of prompt for PARA translations with one demonstration and a text to translate.

that they were qualified to evaluate the translations. All translators were highly proficient in the source language and most of them were native speakers of the target language with some being bilingual.[33] Only one translator reported familiarity with the book, which translation she evaluated. All translators were instructed to evaluate each paragraph in isolation without relying on any prior knowledge about the book and to allow for *all* possible interpretations based on the given part of the source text. They were asked to evaluate five translations first and received feedback on their work before moving forward. Details about the translators are reported

[33]We consider a translator bilingual only if they were raised using both languages; i.e. both can be consider their *native* languages (e.g., *ru-pl* translator was raised in Poland while speaking Russian at home). In the broader sense of this word, all of the translators are bilingual with some of them being trilingual. For the cases where the hired translator was *not* a native speaker of the target language, the annotations were verified by a native speaker of the target language in consultation with the translator.

Book	Lang Pair	Source	Target
An Inventory of Losses	de-ja	Natürlich hatte ich schon davor andere bemerkenswerte Begräbnisstätten besucht: die Toteninsel San Michele etwa, wie sie mit hohen, roten Backsteinmauern aus dem blaugrünen Wasser der Lagune von Venedig emporragt gleich einer uneinnehmbaren Festung, oder das grelle Jahrmarktstreiben des Hollywood Forever Cemetery am alljährlich von der mexikanischen Bevölkerung begangenen Día de los Muertos mit den orange-gelb geschmückten Gräbern und den von der fortgeschrittenen Verwesung auf ewig zum Grinsen verdammten Totenschädeln aus bunt gefärbtem Zucker und Pappmaché. Doch keine hat mich so berührt wie der Friedhof jener Fischersiedlung, in dessen eigentümlichem Grundriss — einer Art Kompromiss aus Kreis und Quadrat ich nichts anderes als ein Sinnbild der ungeheuerlichen Utopie zu erkennen glaubte, die ich dort verwirklicht sah: mit dem Tod vor Augen zu leben. Lange Zeit war ich überzeugt, an diesem Ort, dessen dänischer Name »kleine Insel« oder »vom Wasser umgeben« bedeutet, sei man dem Leben näher, gerade weil seine Bewohner die Toten wortwörtlich in ihre Mitte geholt hatten, anstatt sie wie sonst in unseren Breitengraden üblich — aus dem Innersten der Gemeinden vor die Stadttore zu verbannen, auch wenn der urbane Raum sich die Gräberstätten durch sein ungehemmtes Anwachsen oft nur wenig später wieder einverleibt hat.	もちろんそれ以前にもいくつか特筆すべき墓所を訪れたことはあった。たとえばヴェネツィアの干潟の青緑色の水中から、赤煉瓦の高い壁に囲まれて難攻不落の要塞のようにそびえたつ死者の島、サン・ミシェル。あるいはメキシコ系住民が毎年にぎやかに死者の日を祝う、ハリウッド・フォーエバー墓地。墓はオレンジと黄色の花で飾られ、カラフルな砂糖菓子や張り子細工の頭蓋骨は、腐敗が進んで永遠の笑顔を浮かべているようだ。けれども、この漁師町の墓地ほどに私の心を動かす墓所はなかった。まるで円と四角の間の妥協のようなその独特の輪郭に、私はまさにユートピアの象徴を見たように思った。死を目の前にしつつ生きるというユートピアが、そこに実現されていた。長いこと私は確信していた。デンマーク語で「小さな島」とか「水に囲まれた」という意味の名前を持つこの場所に住む人々は、同じくらいの緯度の国々で通常行われているように、共同体の内部から市門の外へと死者たちを追放する代わりに、死者たちを文字通り町の中心に迎え入れた。だからこそ、より生に近いのだと。もっとも都市空間もまた人口膨張のために、ほとなくして墓地をふたたび内部へと取り込まざるを得なくなるのだけれど。
A Children's Bible	en-pl	The lady urinated. "Oh, poor old thing, she has a nervous bladder!" exclaimed someone's chubby mother. "Is that a Persian rug?" Whose mother was it? Unclear. No one would cop to it, of course. We canceled the performance. "Admit it, that was your mother," said a kid named Rafe to a kid named Sukey, when the parents had filed out. Some of their goblets, highball glasses, and beer bottles were completely empty. Drained. Those parents were in a hurry, then. "No way," said Sukey firmly, and shook her head. "Then who is your mother? The one with the big ass? Or the one with the clubfoot?" "Neither," said Sukey. "So fuck you."	Dama się posikała. – Och, biedactwo, ma wrażliwy pęcherz! – wykrzyknęła czyjaś pulchna matka. – Zaraz, to perski dywan? Czyją matką była? Nie wiadomo. Oczywiście nikt nie chciał się przyznać. Odwołaliśmy przedstawienie. – No dawaj, to twoja – powiedział chłopiec imieniem Rafe do dziewczynki imieniem Sukey, kiedy rodzice sobie poszli. Zostawili po sobie kieliszki, wysokie szklanki i butelki po piwie. Niektóre były zupełnie puste. Do ostatniej kropelki. Tym z rodziców się zatem spieszyło. – W życiu – odparła Sukey stanowczo i pokręciła głową. – To która? Ta z wielkim dupskiem? Czy ze szpotawą stopą? – Ani jedna, ani druga. Spierdalaj.

Table 8: Examples of aligned reference source and target paragraphs from our dataset, including both a narrative (*An Inventory of Losses*) and a dialogue (*A Children's Biblie*). Our PARA approach takes as input the entire source paragraph and outputs a paragraph-level translation.

LANG PAIR	TITLE	AUTHOR	TRANSLATOR(S)	YEAR PUBLISHED	
				TRANSLATION	ORIGINAL
ja-pl	Norwegian Wood	Haruki Murakami	Dorota Marczewska & Anna Zielińska-Elliott	1987	2006
de-pl	The Trial	Franz Kafka	Jakub Ekier	1925	2008
fr-pl	Les Miserables	Victor Hugo	Krystyna Byczewska	1862	1966
fr-pl	The Little Prince	Antoine de Saint-Exupéry	Jan Szwykowski	1862	1967
en-pl	The Valley of Fear	Arthur Conan Doyle	Tadeusz Evert	1915	1927
ru-pl	War and Peace	Leo Tolstoy	Andrzej Stawar	1869	1958
cs-pl	War with Newts	Karel Čapek	Jadwiga Bułakowska	1936	1949
pl-ja	Solaris	Stanisław Lem	Mitsuyoshi Numano	1961	2004
ru-ja	Anna Karenina	Leo Tolstoy	Hakuyō Nakamura	1878	2004
de-ja	Der Steppenwolf	Hermann Hesse	Fujio Nagano	1927	2000
fr-ja	Around the World in 80 Days	Jules Verne	Yū Takano	1873	2009
en-ja	Animal Farm	George Orwell	Eitarō Sayama	1945	1998
zh-ja	Medicine	Lu Xun	Kōbai Inoue	1919	1919
zh-ja	The True Story of Ah Q	Lu Xun	Kōbai Inoue	1921	1923
zh-ja	Diary of a Madman	Lu Xun	Kōbai Inoue	1921	1923
ru-en	Confession	Leo Tolstoy	Peter Carson	1882	2013
zh-en	The Day the Sun Died	Yan Lianke	Carlos Rojas	2015	2018
ja-en	Kokoro	Natsume Sōseki	Edwin McClelan	1914	1957
ja-en	Kokoro	Natsume Sōseki	Meredith McKinney	1914	2010
de-en	Venus in Furs	Ritter von Leopold Sacher-Masoch	Fernanda Savage	1870	*unclear*
fr-en	The Debacle	Émile Zola	Leonard Tancock	1870	1972

Table 9: List of novels employed in the prompts.

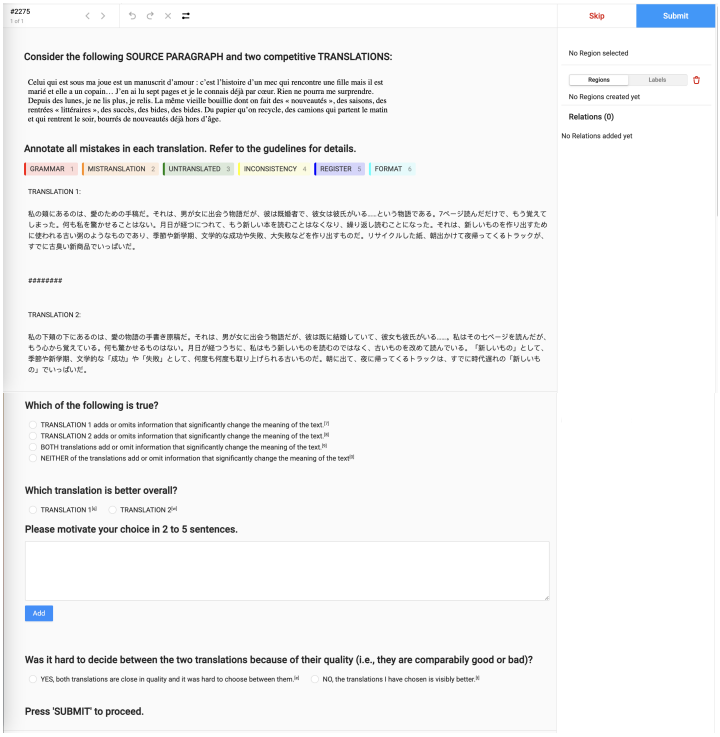


Figure 11: The annotation interface used for the error annotation task.

LANG PAIR	NATIVE LANG	BOOK FAMILIARITY	GENDER
zh-en	Chinese	✗	Male
ja-en	English	✗	Male
de-en	Polish/English	✗	Female
fr-en	English	✗	Female
ru-en	Russian	✗	Female
pl-en	Polish/English	✗	Female
en-ja	Japanese	✗	Female
fr-ja	Japanese	✗	Male
de-ja	Japanese	✗	Female
pl-ja	Polish (author)	✗	Female
ru-ja	Japanese	✗	Male
zh-ja	Japanese	✗	Male
de-pl	Polish/English	✗	Female
en-pl	Polish (author)	✗	Female
ru-pl	Polish/Russian	✗	Female
cs-pl	Czech	✗	Male
ja-pl	Polish (author)	✗	Female
fr-pl	Polish	✓	Female

Table 10: Details about the translators hired for the current annotation study. We note whether the translator was familiar with the source text prior to the evaluation task (*Book Familiarity*).

in Table 10.[34]

E Results

In this section of the appendix, we provide more detailed analysis of the results of the human evaluation. We start with providing more details about the GTR vs PARA evaluation. Next, we include an in-depth discussion of the context-related errors in SENT which were corrected in the PARA translations. Finally, we include some comments from the translators. In the next section (§F), we also provide more information about the issues still present in the PARA translations along with the preliminary analysis of paragraph-level translation by GPT-4.

E.1 PARA is clearly better than GTR

PARA translations are overwhelmingly preferred over those from Google Translate (GTR), with an 82.8% preference rate (p<.001, 95% CI [0.765, 0.880]). Even after removing the "unsure" votes, the preference for PARA remains significant at 88.0% (p<.001, 95% CI [0.812, 0.930]). In the *fr-ja*, *pl-ja*, *zh-ja*, and *cs-pl* language pairs, PARA received *all* of the ten votes over GTR. Part of this advantage may be attributed to GTR sometimes using English as a pivot language, which can result in information loss. Our Czech translator observed that mistakes in GTR translations suggest the text was first translated into English.[35]

[34]Three language pairs (*pl-ja*, *en-pl*, *ja-pl*) were annotated by the first author of this paper.

[35]For the *cs-pl* language pair, we separately annotated mistranslations arising from pivot translation. These errors accounted for over 50% of all mistranslations in that language pair. The elimination of the need for parallel data may therefore be beneficial for translating between lower-resource

Overall, GTR translations result in 57.7% more mistranslations, 37.3% more grammatical errors, over twice as many inconsistency errors, and ten times more register errors (see Table 2). Additionally, GTR produced 125 format errors while PARA produced perfect outputs in this regard. Finally, it is worth noting that GTR left fewer words untranslated, though this is inflated by the fact that in one German text, the word "Bauer" ("farmer") was untranslated 14 times in the PARA translation.

E.2 Context-related errors

Here we present examples of context-related issues present in SENT while correctly translated by PARA.[36]

Pronouns: Unsurprisingly, the absence of discourse context results in the incorrect translation of pronouns. Consider the following example, with English glosses of important words provided in [brackets]:

(2) И ветер [*wind*] то начинал шуметь в голых деревьях, то замолкал, так же как и я прислушиваясь к течению ночи. Но он [*he*] не уходил, он [*he*] был здесь.
—RUSSIAN SOURCE (from *The Story of a Life*)

a. The wind would start to rustle in the bare trees and then fall silent, just as I listened to the flow of the night. But he didn't leave, he was here.
—GPT-3.5 SENT (ENGLISH)

b. The wind would start to rustle in the bare trees, then die down, just like me, listening to the flow of the night. But it didn't go away, it was still here.
—GPT-3.5 PARA (ENGLISH)

In Russian, nouns have grammatical gender. "Wind" in the first sentence of the source text is a masculine noun, so it is later referred to as "he" in (2). Without access to the context, the SENT model incorrectly translates it as "he" into English (2a), while the PARA translation correctly modifies the pronoun to "it" (2b).

When translating from Russian into Polish, another language with grammatical gender, we observe issues when the gender of Russian and Polish nouns differs. Consider the following example:

(3) Романы, как известно, печатались на разной бумаге [*paper*]. И гореть она [*she*] может поразному.
—RUSSIAN SOURCE (from *Manaraga*)

a. Romany, jak wiadomo, drukowano na różnym papierze [*paper*]. I może ona [*she*] tęsknić na różne sposoby.
—GPT-3.5 SENT (POLISH)

b. Jak wiadomo, powieści drukowano na różnym papierze [*paper*]. I może on [*he*] palić się na różne sposoby.
—GPT-3.5 PARA (POLISH)

Although both Russian and Polish nouns possess grammatical gender, "Paper" in (3) is feminine in Russian and referred to as "she," whereas it is a masculine noun in Polish and should be referred to as "he," as in (3b). The absence of context in SENT leads to an incorrect translation in (3a).

Cultural nuances: Assigning appropriate pronouns without context becomes even more challenging when translating from languages like Japanese, in which speakers frequently refer to the listener (or themselves) in the third person rather than using second-person personal pronouns such as "you" in English. Consider the following example:

(4) 「気が付かなくてすみません」
「いやいや、(...)。古倉さんは毎日勤務なのに手を抜かないからねー！」
[lit. *Ms./Mrs./Mr. Furukura works every day*]
—JAPANESE SOURCE (from *Convenience Store Woman*)

a. "I'm sorry I didn't notice."
"No, no, (...). Furukura-san works hard every day without taking any shortcuts!"
—GPT-3.5 SENT (ENGLISH)

b. "I'm sorry I didn't notice."
"No, no, (...). You work every day, but you never slack off!"
—GPT-3.5 PARA (ENGLISH)

From the context of this conversation, a Japanese listener can easily infer that "Furukura-san" or "Miss Furukura"[37] in the last source sentence (4) is used instead of the second-person "you" as per Japanese convention. Translating this sentence without context into English, a language in which third-person reference is not common,[38] results in a confusing translation (4a) that implies that the speaker refers to some other "Furukura" rather than their listener. However, when translating the sentence in context, the model correctly changes "Furukura" into "you" (4b), which makes it clear whom the speaker refers to in English.

languages where sufficient parallel data is often unavailable necessitating the pivot translation.

[36]Note that PARA also suffers from context-related issues. However, at a much lesser extent than SENT.

[37]Note that the gender of neither character is apparent from the fragment alone.

[38]While third-person reference can be used in English, it is only used in rare circumstances e.g. when addressing children.

Language Pair	Sent	Para	Para_Sent	Para	GTr	Para
Russian - English	0	10	5	5	4	6
Chinese - English	1	9	3	7	3	7
Polish - English	4	6	4	6	1	9
French - English	5	5	4	6	2	8
Japanese - English	1	9	2	8	1	9
German - English	5	5	3	7	4	6
Total	16	**44**	21	**39**	15	**45**
Percentage	26.67%	**73.33%**	35.00%	**65.00%**	25.00%	**75.00%**
German - Japanese	6	4	3	7	1	9
Russian - Japanese	4	6	4	6	2	8
French - Japanese	2	8	1	9	0	10
Polish - Japanese	2	8	4	6	0	10
English - Japanese	3	7	2	8	1	9
Chinese - Japanese	4	6	4	6	0	10
Total	21	**39**	18	**42**	4	**56**
Percentage	35.00%	**65.00%**	30.00%	**70.00%**	6.67%	**93.33%**
English - Polish	0	10	3	7	4	6
Japanese - Polish	3	7	5	5	1	9
French - Polish	4	6	4	6	2	8
Czech - Polish	3	7	2	8	0	10
Russian - Polish	1	9	4	6	3	7
German - Polish	3	7	3	7	1	9
Total	14	**46**	21	**39**	11	**49**
Percentage	23.33%	**76.67%**	35.00%	**65.00%**	18.33%	**81.67%**
TOTAL	51	**129**	60	**120**	30	**150**
PERCENTAGE	28.33%	**71.67%**	33.33%	**66.67%**	16.67%	**83.33%**

Table 11: The number of votes for Sent vs Para, Para_Sent vs Para, and GTr vs Para in human evaluation by the language pair. The winning counts are highlighted in purple.

Ellipsis: Another example where context helps is the translation of elliptical constructions. Consider the following example:

(5) „Ne, ted' uděláš nádobí!“ [*(you) will do the dishes!*]
„Neudělám!“ [*(I) won't do!*]
„Uděláš!“ [*(You) will do!*]
—Czech Source (from *Crows*)

a. — Nie, teraz zrobisz zmywanie! [*(you) will do the washing*]
— Nie zrobię! [*(I) won't do!*]
— Zrobisz to! [*(You) will do it!*]
—GPT-3.5 Sent (Polish)

b. — Nie, teraz umyjesz naczynia [*(You) will wash the dishes*]!
— Nie umyję [*(I) won't wash*]!
— Umyjesz [*(You) will wash*]!
—GPT-3.5 Para (Polish)

Czech uses the same collocation as English, "*do* the dishes" (5), which is invalid in Polish. Hence, the ellipses in the last two sentences in (5) require broader context to be translated correctly. Para does it properly, translating both as "wash" (5b), while Sent unsurprisingly fails to choose the correct collocation (5a).

Subject ellipsis: Similarly, context may be needed to attribute a state or an action to the correct character due to the subject ellipsis. This is an obvious issue for languages like Japanese, which tend to omit the subject of the sentence and do not encode any relevant information in the verb form, but it can also arise in English. Consider the following example:

(6) When we were done, the lipstick went back into some mother's Fendi handbag. We watched her apply it, unaware.
—English Source (from *A Children's Bible*)

a. Gdy skończyliśmy, szminka wróciła do jakiejś torebki Fendi należącej do matki. Patrzyliśmy, jak to robi, nieświadomi [*unaware (we)*] tego.
—GPT-3.5 Sent (Polish)

b. Kiedy skończyliśmy, szminka wróciła do torebki Fendi jakiejś matki. Patrzyliśmy, jak ją nakłada, nieświadoma [*unaware (she)*] naszych działań.
—GPT-3.5 Para (Polish)

From the second sentence alone it is not clear who is "unaware" (6) – the mother or the "we" (referring to children) watching her. Only from the

broader context can we confidently deduce that it is in fact the mother, not the children, who is "unaware." PARA (6b) correctly attributes the state of being "unaware" to the mother, which is exhibited by its usage of the singular feminine form of the adjective. In contrast, SENT (6a) mistranslates it using the plural masculine form of the adjective "unaware," which implies that it refers to "we" rather than the "mother."

Consistency: Context is sometimes critical for preserving the overall consistency of the text. The simplest cases include referring to the same entity – a place or a person – in the same way. More interesting cases pertain to style and can enhance the reader's experience. Consider the following example:

(7) Alles zu vergessen, ist gewiss schlimm [*bad*]. Noch schlimmer [*worse*] ist, nichts zu vergessen (...).
—GERMAN SOURCE (from *An Inventory of Losses*)

a. すべてを忘れることは確かに悲惨な[*tragic*]ことです。さらに悪い[*worse*]のは、何も忘れないことです。
—GPT-3.5 SENT (JAPANESE)

b. すべてを忘れることは確かに悪い[*bad*]ことです。もっと悪い[*worse*]ことは、何も忘れないことです。
—GPT-3.5 PARA (JAPANESE)

The German source in (7) translates into English as "To forget everything is *bad*, certainly. *Worse* still is to forget nothing."[39] It is arguably important for the translation to repeat the same word which is an equivalent of the German "schlimm" ("bad"). PARA does it well, translating both as 悪い "warui," or "bad" (7b), in the exact same way as the human Japanese translator. SENT, on the other hand, uses two different words, "tragic" and "bad" (7a), which while technically correct omits the intentional repetition that is meant to introduce an unexpected conclusion.

Polysemy: The absence of context makes it difficult to interpret words or expressions that have multiple meanings in the source language. Consider the following example:

(8) Все прошло хорошо. Книга прочитана идеально – не быстро и не медленно, минимум дыма. Классика. Я был в форме [*in shape*].
—RUSSIAN SOURCE (from *Maranaga*)

a. Wszystko poszło dobrze. Książka została przeczytana idealnie – nie szybko i nie wolno, minimalna ilość dymu. Klasyka. Byłem w mundurze [*in uniform*].
—GPT-3.5 SENT (POLISH)

b. Wszystko poszło dobrze. Książka przeczytana idealnie – nie szybko i nie wolno, minimalna ilość dymu. Klasyka. Byłem w formie [*in shape*].
—GPT-3.5 PARA (POLISH)

The ambiguity stems here from multiple meanings of the Russian noun форма "forma" (8), which can mean either "shape" or "uniform." Since one can be "in shape" as well as "in a uniform", it is unclear from the sentence alone which meaning was intended by the author. From the preceding context, it is clear that "everything went well" for the narrator, who mastered the art of "book'n'grill," a unique form of expression exclusive to this fictional world. Based on this, we can infer that in this instance, the term "forma" signifies "shape," as in (8b), rather than "uniform," as in (8a).

Appropriateness: Finally, context may help to choose the more appropriate equivalent for the given situation. Consider the following example:

(9) 「あー、あと煙草の５番を一つ」
「かしこまりました」 [lit. *(I) understood*]
—JAPANESE SOURCE (from *Convenience Store Woman*)

a. "Ah, and one pack of cigarettes, number five."
"Understood."
—GPT-3.5 SENT (ENGLISH)

b. "Ah, and one pack of cigarettes, number five."
"Right away."
—GPT-3.5 PARA (ENGLISH)

The conversation above is between a clerk and a customer. The Japanese expression かしこまりました "kashikomarimashita" (9) is an honorific that literally means "understood." However, when choosing the best equivalent, the translator needs to consider the situation at hand to best reflect its meaning in the target language. "Understood" in SENT (9a) is technically correct, but it is an unfortunate word choice for the clerk to employ. On the other hand, "right away" in PARA (9b) fits much better in the context of this conversation. Had this been a series of commands (e.g., in a military context) "understood" would be the more favorable option.

E.3 What do translators think about PARA?

To wrap up this section, we provide a qualitative analysis of the free-form comments written by translators to justify their preference judgments. Overall, the translators praise PARA for its *more skillful use of rhetoric devices*, and *surpas[ing]*

[39] Excerpt taken from the official English translation by Jakie Smith (2020).

SENT *as a literary rendition*. They also mention that PARA *uses more of a poetic license but this makes it stylistically much smoother* than SENT. Furthermore, translators state that PARA *clearly better reflects the content and style of the original* when compared to GTR, and that it *stays consistent within the paragraph*. Inevitably, translations are not flawless, and there are instances where both compared systems fall short, as highlighted by one of the translators when assessing PARA against SENT: *Nightmare, a mistake upon mistake (...) Despite all these mistakes, I can understand the* [PARA] *translation better but they are equally miserable.*

F Limitations

In this section of the appendix, we delve deeper into the unresolved issues in the PARA translations. First, we discuss the omissions present in the translations. Next, we highlight some mistranslations that persist in the PARA translations. To conclude, we briefly discuss our initial experiments utilizing GPT-4 for paragraph-level translation.

Omissions: One thing we ought to discuss is the omission issue. Upon examining translations and annotator feedback, we observe that PARA occasionally omits details, which are crucial to the storyline. Preliminary investigation indicates that PARA translations are more prone to omissions compared to SENT and GTR. Although PARA_SENT appears to mitigate this problem to some extent, it still results in a higher number of omissions than the sentence-level approach while at the same time introducing some repetition issues (see Table 12).[40]

Mistranslations: Moreover, PARA still makes a sizeable number of mistranslations and grammatical errors, though fewer than SENT or GTR. We observe that PARA occasionally merges sentences with two distinctive subjects attributing all states and/or actions to one of them. Very rarely, we also find cases where context possibly confuses the model, resulting in an incorrect translation. The following example illustrates this issue:

(10) Le bois du bureau amplifie les battements de mon cœur. Le vieux mobilier Art déco conduit bien les émotions et les fatigues. Ruhlman ? Leleu ? Il [*he*] en a tant vu.

—FRENCH SOURCE (from *Dear Reader*)

a. 机の木材が私の心臓の鼓動を増幅している。古いアール・デコ家具は感情や疲労をうまく導いてくれる。ルールマン？レルー？彼ら [*they*] はそんなに多くを見てきた。

—GPT-3.5 PARA (JAPANESE)

In the French text, the narrator wonders whether the brand of the desk was Ruhlman or Leleu, with both proper nouns possibly referring to a person. In the last sentence, the French text uses "il" or "he" (10), as a desk is a masculine noun in French ("le bureau"). PARA, on the other hand, appears to be confused by the two preceding names and incorrectly translates the singular pronoun as 彼ら, or "they."

Furthermore, we observe (very few) cases where the paragraph-level translation disregards the context. Most representative of this class of errors is when the model struggles to translate from Japanese in cases where the subject is omitted. The following example illustrates this issue:

(11) ミホ [*Miho*] は、今では結婚して地元に中古の一戸建てを買っていて、そこに友達がよく集まっている。明日もアルバイトなので億劫に思う時もあるが、コンビニ以外の世界との唯一の接点であり、同い年の「普通の三十代女性」と交流する貴重な機会なので、ミホの [*Miho's*] 誘いにはなるべく応じるようにしている。

—JAPANESE SOURCE (from *Convenience Store Woman*)

a. Miho [*Miho*] wyszła za mąż i kupiła stary, jednorodzinny dom w swoim rodzinnym mieście. Przychodzą tam często jej znajomi. Mimo że Miho ma [*Miho has*] jutro pracę w konbini, zazwyczaj chętnie odpowiada [*(she) responds*] na jej [*her*] zaproszenia, bo to jedyna okazja, by spotkać się z innymi kobietami w jej [*her*] wieku.

—GPT-3.5 PARA (POLISH)

b. Miho is now married and has bought an old house in her hometown, where her friends often gather. Though she often finds it a chore to work tomorrow, it is her only connection to the world outside the convenience store, and a valuable opportunity to interact with other "normal thirty-something women" her age, so she tries to accept Miho's invitations as often as possible.

—GPT-3.5 PARA (ENGLISH)

Both Polish (11a) and English (11b) translations of the same source text (11) share a common issue. The narrator begins the paragraph by talking about Miho and then proceeds to describe her

[40] Note that although ask the annotators to report both omissions *and* additions, based on their comments and our analysis of the translations, we conclude that omissions are the predominant issue. In version two of our data (currently on https://github.com/marzenakrp/LiteraryTranslation), we further annotate any repetition as a separate type of error (i.e. error with 'repetition' label) rather than counting it as an addition. This annotations resulted in eight repetition errors in PARA_SENT translations.

Language Pair	PARA	SENT	PARA_SENT	GTR
Russian-English	0	0	1	0
Chinese-English	1	0	1	0
Polish-English	0	0	0	0
French-English	1	0	2	0
Japanese-English	2	1	2	3
German-English	0	0	0	0
German-Japanese	8	2	6	8
Russian-Japanese	10	4	6	4
French-Japanese	3	1	4	4
Polish-Japanese	4	1	3	0
English-Japanese	2	2	1	0
Chinese-Japanese	2	0	0	1
English-Polish	0	1	2	0
Japanese-Polish	0	0	1	1
French-Polish	2	2	1	1
Czech-Polish	1	2	1	0
Russian-Polish	1	1	1	0
German-Polish	0	0	0	0
Total	**37**	17	32	22

Table 12: Count of omissions reported by the translators for each translation method.

own (the narrator's) feelings about the situation, although the gender of the narrator is never revealed in the Japanese text. The second sentence should be written from a first-person perspective, particularly since it directly references Miho towards the end (blue text). However, both the Polish and English translations produced by PARA are confused by this: by using the third-person's perspective ("she," "her"), both translations incorrectly imply that Miho is the subject of the second sentence. SENT and GTR translate this passage accurately, albeit with some clumsy phrasing.

GPT-4 does not magically solve all of these issues! Our preliminary experiments indicate that GPT-4 (OpenAI, 2023) sometimes generates better paragraph-level translations than those of GPT-3.5. For instance, it seems to have a better grasp of the inverted word order in German, though no broader conclusions should be made without further testing. Nevertheless, it does not resolve all of the issues discussed in our paper. Mistranslations and grammatical errors are still abundant across many language pairs. GPT-4 produces the following translation when fed the previous example paragraph (11) as input; note that all of the issues still remain:[41]

(12) Miho is now married and has bought a used single-family home in her hometown where her friends often gather. Although she sometimes finds it a drag to work a part-time job the next day, she makes an effort to respond to Miho's invitations because it's a valuable opportunity to interact with "normal" women in their thirties like herself, apart from her convenience store job.

—GPT-4 PARA (ENGLISH)

PARA translations hold the potential to captivate readers, especially if LLMs continue to improve at their current pace. Indeed, some of our translators mentioned that they genuinely enjoyed the task, though integrating these paragraphs into a coherent novel still poses a considerable challenge. With all that said, literary translation involves more than just overall "correctness" or mere entertainment value. A translation that is perfectly "correct" and enjoyable might still fail to convey the author's intentions or meaning skillfully hidden behind a simple phrase. Our *fr-en* translator shares her thoughts on this matter:

[41] Although the given paragraph is already comprehensible for a human reader, we also attempt to enhance the translation by incorporating three additional preceding paragraphs for context. Intriguingly, when provided with this extended context, both GPT-3.5 and GPT-4 generated accurate translations.

System	Comet	Bleurt	BertScore	Comet-QE
Para	**0.785**	**0.485**	**0.840**	**0.038**
Sent	0.779	0.469	0.839	-0.052
Para_Sent	0.780	0.480	0.838	-0.062
GTr	0.735	0.443	0.832	-0.156

Table 13: Results of automatic evaluation. A higher number indicates better scores.

Metric	Acc	τ	Acc (*conf*)	τ (*conf*)
Comet	67.41%	0.348	72.78%	0.456
Comet-QE	64.44%	0.289	70.64%	0.413
Bleurt	61.30%	0.226	66.36%	0.327
BARTScore	58.52%	0.170	63.91%	0.278

Table 14: Correlation of automatic metrics with human judgments from our human evaluation. We evaluate the metrics performance on *all* human judgments as well as on the *subset* of judgments where the translator indicated that the chosen translation was visibly better (*conf*). We report both the percentage of agreement (Acc) and Kendall's Tau (τ). Data reported on v1 of the dataset.

> *Both translations* [Sent and Para] *translate the words without the feeling; the original author's voice is lost.*
>
> —French to English Translator

G Automatic Evaluation

In this section of the appendix, we present the results of automatic evaluation. First, we discuss the scores assigned to the translations by automatic metrics.[42] Then we provide the statistical analysis. Finally, we present the correlation of each metric with human judgments for the 180 paragraphs used in the human evaluation.

Automatic metrics favor Para: We assess the translation from all four systems using the reference-based Comet (Rei et al., 2022), Bleurt (Sellam et al., 2020), and BertScore (Zhang et al., 2020) metrics, as well as the reference-free Comet-QE (Rei et al., 2021)[43] metric. Although these metrics were not explicitly designed for evaluating paragraph-level outputs and their results should be interpreted with caution, they prove more reliable than string-based metrics like Bleu, especially for literary translations (Thai et al., 2022; Karpinska et al., 2022; Gehrmann et al., 2022). Table 13 shows the effectiveness of the Para translation method: a statistical analysis with linear mixed-effects models (Baayen et al., 2008) demonstrates that Para significantly outperforms Sent and GTr based on Comet, Bleurt, and Comet-QE scores (p<.001), and surpasses GTr based on the BertScore results (p<.001). We discuss the details of this statistical analysis in the next section.

Statistical Analysis: We employ the linear-mixed effect models (Baayen et al., 2008) to analyze the scores produced by automatic metrics. We fitted the model in `R` using the `lme4` package (Bates et al., 2015); the p-values were obtained with the `LmerTest` package (Kuznetsova et al., 2017). Linear-mixed effects models contain both *fixed-effects* and *random-effects* (random *intercept* and/or *slope*). The fixed effect here is the translation setup (Para, Sent, Para_Sent, GTr) with the source paragraph being coded as the random effect (random intercept). We inspect the residual plots to ensure that the variance across the fitted range is relatively constant. The results from the fitted model are presented in Table 18 (Bleurt), Table 20 (Comet), Table 22 (Comet-QE), and Table 24 (BertScore).[44]

We further perform a post hoc analysis using the `emmeans` package (Lenth, 2023) to obtain p-values for the pairwise comparison. The results of the post hoc analysis are presented in Table 19 (Bleurt), Table 21 (Comet), Table 23 (Comet-QE), and Table 25 (BertScore).

Correlation with Human Judgements: We investigate the correlation of automatic metrics with human judgments in our evaluation. We consider (1) all the judgments, as well as (2) a subset of all judgments where the annotator stated that they were sure that one translation is *clearly* better than the other. We compute both *accuracy* (i.e., the percentage of cases where the metric agrees with human judgment), and a correlation coefficient Kendall's Tau which is defined as follows:

$$\tau = \frac{\text{Concordant} - \text{Discordant}}{\text{Concordant} + \text{Discordant}}$$

[42] This analysis is done on the entire dataset excluding only the paragraphs which were too long as per each metric's token limit.

[43] We use the newest `wmt22-comet-da` checkpoints for Comet, `Bleurt-20` checkpoints for Bleurt, `wmt20-comet-qe-da` checkpoints for Comet-QE, and the HuggingFace implementation which employs `roberta-large` for BertScore.

[44] It should be noted that, while significant, the analysis is underpowered. It is possible that analyzing more examples would provide a more reliable analysis.

SOURCE	TARGET	PARA	PARA_PIVOT
Czech	*Polish*	11	9
German	*Japanese*	13	7
German	*Polish*	12	8
French	*Japanese*	9	11
French	*Polish*	11	9
Japanese	*Polish*	10	10
Polish	*Japanese*	3	17
Russian	*Japanese*	10	10
Russian	*Polish*	8	12
Chinese	*Japanese*	9	11
	TOTAL	*96*	*104*

Table 15: The results of pairwise comparison for the paragraph-level translations with (PARA_PIVOT) and without (PARA) English as a pivot language.

Table 14 shows the correlation of automatic metrics with the human judgments obtained in this study. COMET exhibits the highest agreement with human judgments both in terms of the *accuracy* (64.04% for all data, 72.78% for confident votes only) and Kendall's Tau (0.341 for all data, 0.456 for confident votes only).

H Pivot Pilot

In this section of the appendix, we discuss the results of the preliminary study where we translated the paragraphs using English as a pivot language. Table 15 shows the results of this pilot study. The evaluation was done by the first author on *all* 20 passages for every language pair that did not include translation from or into English, as these do not require any pivoting.[45] A total number of 200 pairs was evaluated employing simple preference judgments.

During the PARA_PIVOT translation process, the model utilized both the source text and its corresponding English translation (`text-davinci-003`, top-p=1.0, temp=0.3). This approach has the potential to mitigate the limitations associated with pivoting translations, where some information may be lost. For example, both Czech and Polish encode the gender information in the past tense form of the verb. English does not, so this information is lost and will most likely result in an erroneous translation. Indeed, we notice that adding the source text helps the model to overcome this shortcoming, however, we do not observe a clear gain from using English as a pivot language. Consider the following example:

(13) Taťka mě zase zmlátil. Byl hrozně naštvanej, protože mamka řvala. On mě vždycky mlátí, když ona řve. Že prej jsem pořádně neudělala nádobí. Ale já vím, že jsem je udělala dobře. I ten hrnec jsem vydrhla pořádně. A ještě jsem to všechno utřela před koupáním. Ale možná jsem nevyždímala hadr.
—CZECH SOURCE (from *Crows*)

a. Tata mnie znowu zbił. **Był wściekły**, bo mama krzyczała. On zawsze mnie bije, kiedy ona krzyczy. Że niby nie **umyłam** dobrze naczyń. Ale ja wiem, że **umyłam** je dobrze. I garnek też dokładnie **wypolerowałam**. I jeszcze wszystko **wytrzepałam** przed kąpielą. Ale może nie **wyciśniłam** ręcznika.
—GPT-3.5 PARA (POLISH)

b. Tata mnie znów uderzył. **Był wściekły**, bo mama krzyczała. On zawsze mnie bije, kiedy ona krzyczy. Że niby nie **umyłam** dobrze naczyń. Ale ja wiem, że **umyłam** je dobrze. Nawet garnek dokładnie **wytrzepałam**. I jeszcze wszystko przed kąpielą **wytrzepałam**. Ale może nie **wyżągnęłam** mopa.
—GPT-3.5 PARA_PIVOT (POLISH)

In each instance, the emphasized verbs could potentially be mistranslated when translated through English as the pivot language, as the speaker's gender information would be lost. For instance, the past tense verb "washed" remains unchanged in English regardless of the gender of the speaker, with such details encoded only in the source (*Czech*) and target (*Polish*) languages. In this case, all verbs have been translated accurately with respect to grammatical gender, implying that incorporating the source language into the pivot pipeline does indeed improve the translation. However, PARA_PIVOT still selects less suitable verbs (highlighted in red) resulting in slightly more errors in this particular paragraph.

The only pair where pivoting seems to help is *pl-ja*. While it is unclear why this happens, it is possible that this outcome is due to the specifics of the Polish novel employed for the translation. *Sword of Destiny* by Andrzej Sapkowski uses a very distinct language with many archaic expressions. It is possible that translating into English, a language the GPT models were trained on, helps the model deal with these difficult phrases.

Since we do not observe any apparent gains from performing the translation via English as a pivot

[45] The author is fluent in English, Japanese, and Polish with a limited proficiency in other source languages.

language (p=0.62, 95% [0.448, 0.591]) and doing so reduces the number of examples one can fit into the prompt, we continue our experiments with a direct translation.

Type	Description	Trg Lang	Para	Sent	Para_Sent	GTr
Context (sentence)	A mistranslation that results most likely from lack of "understanding" the sentence-level context (e.g., translating "guide" as "doradca," or "adviser" instead of "przewodnik," or "guide"). This can include translating a word or a phrase into one that is semantically related but does not convey the intended meaning, or translation which appear to be an outcome of translating a word semantically related to the source word, instead of the source word itself.	Japanese	114	118	107	158
		Polish	64	67	49	82
		English	30	36	44	59
Context (paragraph)	A mistranslation that results from lack of a beyond-sentence context. This include issues such as polysemy, employment of correct pronouns, or translating elliptical expressions.	Japanese	6	36	6	38
		Polish	13	51	15	59
		English	2	25	0	48
Minor Issue	A minor issue which does not significantly affect the text and can be disputable, such as translating "barked" as "howl."	Japanese	34	25	26	16
		Polish	33	26	16	13
		English	18	11	12	9
Surface Similarity	A translation by word which is similar to the correct translation on the surface level, but has a different meaning (e.g., "Wilczak," a Polish surname, instead of "wilczarz," a "wolfhound").	Japanese	8	6	7	2
		Polish	14	13	16	5
		English	5	5	6	2
Word-by-Word	A translation of longer phrase which is overly literal resulting in confusing and incorrect translation.	Japanese	15	52	34	84
		Polish	17	23	18	33
		English	7	13	5	20
Unrelated Word	A translation with unrelated word such as "klnie" ("swear") instead of "zapuka" ("knock") where no apparent semantic relation could be found.	Japanese	3	2	5	4
		Polish	5	14	10	12
		English	1	3	1	2
Subject Changed	Change of subject. In the case of Para, it occurs mostly due to merging two sentences with two distinctive subjects where all states and/or actions are then assigned to one of them.	Japanese	5	2	2	0
		Polish	6	0	5	3
		English	7	2	5	1
Factuality	A translation that results in change in factuality, such as translating affirmative sentence as negation or translating word by its antonym.	Japanese	4	11	5	7
		Polish	0	2	1	3
		English	1	2	1	1
Non-word	A translation by a non-existent (made up) word. Some examples include skillfully constructed words like 火炎棒 which was generated instead of a "torch." While this word does not exist in Japanese (or Chinese) it follows the compositionality rules of these languages and is fully intelligible to a native speaker (火炎 "fire" and 棒 "stick.")	Japanese	1	2	2	0
		Polish	6	8	9	3
		English	0	0	0	0
mood	Change in the grammatical mood with regard to the source text. Note that the sentence here is *still* grammatically correct but does not reflect the meaning intended by the author.	Japanese	4	9	1	3
		Polish	1	3	4	2
		English	0	0	0	0
Unnecessary Translation	A translation of text which should be left untranslated such as some proper names.	Japanese	0	0	0	0
		Polish	0	3	0	2
		English	1	1	1	1
Language Mismatch	A translation into a language different than the target language (e.g., Chinese instead of Japanese). Note that leaving the word in the source language classifies as an "untranslated" error.	Japanese	2	3	3	2
		Polish	2	0	2	0
		English	0	0	0	0
Number/Time	A translation which changes number or time expression, such as translating 1h15min as 1h30min. Note that these rarely affect the overall meaning of the text. We have not observe cases where this would be a critical issue.	Japanese	3	2	4	3
		Polish	0	0	0	0
		English	5	2	1	3
Pivot Translation (Czech)	A mistranslation that stems from pivoting on English (annotated for cs-pl language pair).	Polish	0	0	0	43
other	Other issues which do not fit into any of the above.	Japanese	24	26	27	17
		Polish	9	14	10	13
		English	10	4	5	4
		Total (*Japanese*)	223	294	229	334
		Total (*Polish*)	170	224	155	273
		Total (*English*)	87	104	81	150
		Total (*All*)	**480**	**622**	**465**	**757**

Table 16: Classification of mistranslation errors for each system grouped by the target language. The manual classification was performed on the v1 of the annotated dataset.

TRG LANG	TYPE	SUBTYPE	PARA	SENTS	PARA_SENTS	GTR
JAPANESE	PARTICLE	wrong or missing	21	22	13	12
	ADJECTIVE	wrong continuative	0	2	3	0
		other	0	0	2	0
	VERB	tense	3	7	1	14
		mood	2	1	4	5
		finite/non-finite	5	2	1	3
		other	2	5	6	0
	ORDER	wrong order	1	6	1	16
	OTHER		8	5	6	13
	TOTAL		42	50	37	63
POLISH	ADJECTIVE	gender	7	14	8	4
		case	2	1	1	0
		other	1	1	1	1
	NOUN	case	9	13	9	1
		other	3	3	3	2
	PRONOUN	omitted or wrong	5	8	3	2
		case or gender	1	6	4	5
	VERB	aspect	1	5	1	12
		person or gender	2	8	5	2
		conjugation	1	0	7	3
		other	2	4	1	13
	PREPOSITION	omitted or wrong	14	15	15	4
	NUMERAL	case or gender	2	1	0	1
	ORDER	wrong order	2	4	2	4
	OTHER		3	3	4	5
	TOTAL		55	86	64	59
ENGLISH	ARTICLE	omitted or wrong	1	9	2	8
	PREPOSITION	omitted or wrong	3	7	3	5
	OTHER		1	4	4	5
	TOTAL		5	20	9	18

Table 17: Categorization of grammar errors in each translation configuration, grouped by the target language. The manual classification was performed on the v1 of the annotated dataset.

	BLEURT		
Predictors	Estimates	CI	*p*-value
(Intercept)	0.48	0.47–0.50	<**0.001**
PARA_SENT	-0.00	-0.01–0.00	0.130
SENT	-0.02	-0.02–(-0.01)	<**0.001**
GTR	-0.04	-0.05–(-0.04)	<**0.001**

Table 18: Results of linear-mixed effects models analysis for BLEURT scores.

	BLEURT				
Contrast	Estimate	SE	df	*t*-ratio	*p*-value
PARA - PARA_SENT	0.00477	0.00315	1074	1.515	0.780
PARA - SENT	0.01641	0.00315	1074	5.215	<**0.001**
PARA - GTR	0.04155	0.00315	1074	13.205	<**0.001**
PARA_SENT - SENT	0.01164	0.00315	1074	3.700	**0.001**
PARA_SENT - GTR	0.03678	0.00315	1074	11.690	<**0.001**
SENT - GTR	0.02514	0.00315	1074	7.990	<**0.001**

Table 19: Result of post hoc analysis with *emmeans* package for BLEURT.

	COMET		
Predictors	Estimates	CI	*p*-value
(Intercept)	0.79	0.77–0.80	<**0.001**
PARA_SENT	-0.01	-0.01–(-0.00)	**0.019**
SENT	-0.01	-0.01–(-0.00)	**0.004**
GTR	-0.05	-0.05–(-0.05)	<**0.001**

Table 20: Results of linear-mixed effects models analysis for COMET scores.

	COMET				
Contrast	Estimate	SE	df	*t*-ratio	*p*-value
PARA - PARA_SENT	0.00563	0.00239	1074	2.356	0.112
PARA - SENT	0.00691	0.00239	1074	2.893	**0.023**
PARA - GTR	0.04998	0.00239	1074	20.928	**<.001**
PARA_SENT - SENT	0.00128	0.00239	1074	0.536	1.000
PARA_SENT - GTR	0.04435	0.00239	1074	18.571	**<.001**
SENT - GTR	0.04307	0.00239	1074	18.035	**<.001**

Table 21: Result of post hoc analysis with *emmeans* package for COMET.

	COMET-QE		
Predictors	Estimates	CI	*p*-value
(Intercept)	-0.04	-0.06 – -0.01	**0.004**
PARA_SENT	-0.01	-0.03 – -0.00	**0.026**
SENT	-0.02	-0.04 – -0.01	**<0.001**
GTR	-0.12	-0.13 – -0.11	**<0.001**

Table 22: Results of linear-mixed effects models analysis for COMET-QE scores.

			COMET-QE		
Contrast	Estimate	SE	df	*t*-ratio	*p*-value
PARA - PARA_SENT	0.01464	0.00655	1074	2.235	0.154
PARA - SENT	0.02376	0.00655	1074	3.628	**0.002**
PARA - GTR	0.11848	0.00655	1074	18.092	**<.001**
PARA_SENT - SENT	0.00912	0.00655	1074	1.392	0.9844
PARA_SENT - GTR	0.10384	0.00655	1074	15.857	**<.001**
SENT - GTR	0.09472	0.00655	1074	14.464	**<.001**

Table 23: Result of post hoc analysis with *emmeans* package for COMET-QE.

	BERTSCORE		
Predictors	Estimates	CI	*p*-value
(Intercept)	0.84	0.83–0.85	<**0.001**
PARA_SENT	-0.00	-0.00–0.00	**0.037**
SENT	-0.00	-0.00–0.00	0.522
GTR	-0.01	-0.01–0.01	<**0.001**

Table 24: Results of linear-mixed effects models analysis for BERTSCORE scores.

			BERTSCORE		
Contrast	Estimate	SE	df	*t*-ratio	*p*-value
PARA - PARA_SENT	0.002422	0.00116	1074	2.082	0.225
PARA - SENT	0.000745	0.00116	1074	0.640	1.000
PARA - GTR	0.007508	0.00116	1074	6.454	<**0.001**
PARA_SENT - SENT	-0.001678	0.00116	1074	-1.442	0.897
PARA_SENT - GTR	0.005086	0.00116	1074	4.372	<**0.001**
SENT - GTR	0.006763	0.00116	1074	5.814	<**0.001**

Table 25: Result of post hoc analysis with *emmeans* package for BERTSCORE.

Identifying Context-Dependent Translations for Evaluation Set Production

Rachel Wicks[1,2] and **Matt Post**[1–3]
[1]Human Language Technology Center of Excellence, Johns Hopkins University
[2]Center of Language and Speech Processing, Johns Hopkins University
[3]Microsoft
rewicks@jhu.edu, mattpost@microsoft.com

Abstract

A major impediment to the transition to context-aware machine translation is the absence of good evaluation metrics and test sets. Sentences that require context to be translated correctly are rare in test sets, reducing the utility of standard corpus-level metrics such as COMET or BLEU. On the other hand, datasets that annotate such sentences are also rare, small in scale, and available for only a few languages. To address this, we modernize, generalize, and extend previous annotation pipelines to produce CTXPRO, a tool that identifies subsets of parallel documents containing sentences that require context to correctly translate five phenomena: gender, formality, and animacy for pronouns, verb phrase ellipsis, and ambiguous noun inflections. The input to the pipeline is a set of handcrafted, per-language, linguistically-informed rules that select contextual sentence pairs using coreference, part-of-speech, and morphological features provided by state-of-the-art tools. We apply this pipeline to seven languages pairs (EN into and out-of DE, ES, FR, IT, PL, PT, and RU) and two datasets (OpenSubtitles and WMT test sets), and validate its performance using both overlap with previous work and its ability to discriminate a contextual MT system from a sentence-based one. We release the CTXPRO pipeline and data as open source.[1]

1 Introduction

Neural machine translation (NMT) systems can produce high-quality, fluent output which are nearly indistinguishable from human translations, when evaluated at the sentence level. This human-level parity has been shown to disappear, however, when evaluated in context (Läubli et al., 2018; Toral et al., 2018). This is unsurprising, because sentences are nearly always written by humans in some contextual setting, and are translated by translators in the same fashion. Dismissing this context may create ambiguities that do not exist in the document as a whole, and in some cases, may make it impossible to correctly interpret the sentence.

	GENDER	AUXILIARY	INFLECTION	FORMALITY	ANIMACY	LANGS
Müller et al.	✓					de
Lopes et al.	✓					fr
Voita et al.	✓	✓	✓	✓		ru
Nadejde et al.				*		de, es, fr, hi, it, ja
Currey et al.	†					ar, fr, de, hi, it, pt, ru, es
This work	✓	✓	✓	✓	✓	de, fr, ru, pl, pt, it, es

Table 1: This work expands evaluation set coverage to new document phenomena and languages. (*) Note that Nadejde et al. (2022) does not include contextual information. (†) Currey et al. (2022) focuses on natural, rather than grammatical, gender.

Translation to another language must address ambiguities where the semantic or grammatical granularity of two sentences is imbalanced or mismatched. Probably the most widely-known of these is grammatical gender, i.e., when translating referential pronouns from a grammatically non-gendered language to a gendered one. For example, when translating from English to French, the pronoun *it* must be translated to *il* or *elle* depending on the grammatical gender of the antecedent noun, which may not be available in the same sentence.

The obvious path forward in addressing these issues is to move to contextual machine translation, in which sentences are no longer translated in isolation but with their source-side context. Recent work has shown that transformers (Vaswani et al., 2017) are capable of handling longer sequences and improving performance on context-based evaluation (Sun et al., 2022; Post and Junczys-Dowmunt, 2023). However, general contextual translation has

[1]https://github.com/rewicks/ctxpro

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 452–467
December 6–7, 2023.

	English	Target
AUXILIARY	I just figured you need to know. *And now you **do.***	(fr) Je pensais que tu méritais de savoir. *Et maintenant tu **sais.***
	I can't lose my voice. *You **won't.***	(pl) Nie mogę stracić głosu. *Nie **stracisz.***
INFLECTION	Mostly work with the Knicks right now. *And other **athletes.***	(ru) В основном работаю с "Никс". И с другими спортсменами.
GENDER	You think migraines are a sign of weakness, don't want anyone to know. *I used to get **them**, too.*	(it) Lei pensa che le emicranie siano segno di debolezza, e non vuole che si sappia. ***Le** prendevo anch'io.*
	This pain? *I long for **it.***	(pt) A dor? *Anseio por **ela.***
ANIMACY	Et il y a eu cette rose aussi pour toi. *Tu sais, **elle** se distingue des autres.*	(en) Also, uh, this rose came for you. *You know, **it** stands out in front of all the others.*
	La felicidad es un mito. *Y vale la pena luchar por **ella.***	(en) Happiness is a myth. *And **it's** worth fighting for.*
FORMALITY	*We'll call **you** if something happens, huh?*	(de) *Wir rufen **euch** an, wenn etwas passiert.*
	*Well, uh, I was an obstetrician before, and I most definitely owe **you.***	(es) *Bueno, era obstetra antes, y definitivamente se **los** debo.*

Table 2: An example of the extracted ambiguities with their preceding contexts for each language pair. The ambiguous sentence is denoted in *italics* and the ambiguous word is **bolded**. Note the dialectal use of the "usted" accusative form "los". Language denoted in parentheses.

a number of obstacles, foremost is the lack of available evaluation resources. There are essentially two kinds of contextual evaluations: general metrics, which can theoretically be applied to any test set, and fixed test sets. There is relatively little work in the former setting (Vernikos et al., 2022; Jiang et al., 2022), and while they correlate with human judgments, they have not been proven capable of discriminating sentence-based from known-high-quality contextual systems. For the latter, a number of high-quality evaluation sets exist (Müller et al., 2018; Lopes et al., 2020; Bawden et al., 2018; Voita et al., 2019, Table 1), but they are limited both in language coverage and scope of phenomena.

In this work, we address this lack of evaluation data by extending coverage of existing datasets to more languages and contextual phenomena. We:

- develop a pipeline that makes use of broad-language-coverage annotation tools and hand-developed rules to identify context-based phenomena in any test set;
- construct rules for five context-based phenomena (§ 2) and seven language pairs (§ 3): DE, ES, FR, IT, PL, PT, and RU with EN; and
- apply this toolchain to multiple datasets.

We show that this dataset, called CTXPRO, is capable of discriminating high-quality contextual systems from sentence-level ones.

2 Contextual phenomena

A number of context-based phenomena which create ambiguities are common. We display some examples in Table 2. Humans easily handle these ambiguities during translation, which nearly always takes place in context, so a machine translation system which ignores these issues will never reach human-level parity. Some, such as lexical cohesion or fluency, are hard to quantify, while others, for example pronoun translation accuracy or word sense disambiguation, are easier. These phenomena all present difficulties and even impossibilities to systems that translate sentences in isolation. Our goal is to identify as many of these phenomena we can in a general way, such that we can create a general pipeline for isolating them, that can be reliably applied to any test set.

We describe each phenomena for comprehension and then provide our extraction methodology in order to identify when these ambiguities arise.

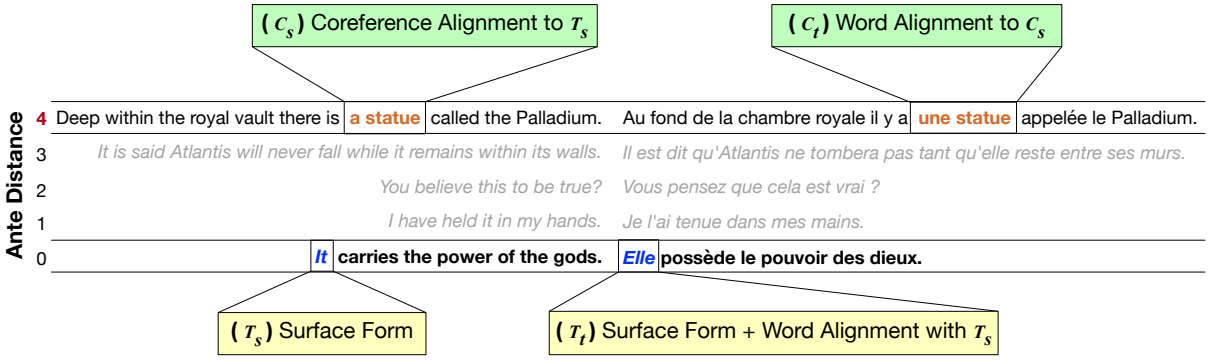


Figure 1: A diagram showing how the four key words for GENDER identification are identified. The antecedent distance is determined by what sentence C_e is found in. In order to be considered, T_e, T_t, C_e, C_t would also have to pass morphological feature tests similar to those shown in Table 3.

2.1 Anaphoric pronouns

Pronouns are a general descriptor that function as a placeholder for a noun phrase, providing the speaker with a more succinct form instead of repeatedly identifying an established referent.

In grammatical contexts, anaphora refers to the use of a pronoun to refer to a previously mentioned word or entity. Pronouns for which the referent noun can be found in preceding contexts are called *anaphora*; in contrast, *cataphora* denotes situations where the referent noun follows the pronoun. We do not consider cataphora in this paper.

2.1.1 Gender

Languages with gendered nouns require agreement with the appropriate gendered pronoun. English, which makes no such distinction for inanimate objects, will use the pronoun "it." In order to correctly translate "it" into Spanish, it is necessary to know what "it" refers to. If "it" refers to a school, it would be translated differently (*una* escuela) than if it refers to a heart (*un* corazón).

Apart from a few exceptions, English does not make use of grammatical gender. Machine translation often centers around translating either into or out of English with most of the paired languages expressing genders (masculine, feminine, and neuter), so there is a clear need to evaluate the translation of gender. Further, removing English from the equation does not resolve the problem. Gender assignment of inanimate objects is arbitrary which means that translating between two gendered languages is non-trivial. In extreme cases, a language may exhibit *"noun classes"* which behave similarly to gender, but may correlate more heavily with meaning. A noun in Swahili is not grouped via an arbitrary *gender* assignment, but is instead somewhat assigned to groups based on other labels such as *animacy, items, plants*, or *tools*. These classes affect morphological agreement in ways that English does not express. In any case, translating a pronoun that refers to a previously mentioned noun requires resolving this coreference in order to correctly generate the new pronoun.

2.1.2 Animacy

Humans and animals are often treated differently grammatically than inanimate objects. As stated, English makes no gender distinction for inanimate objects, though it does have gendered pronouns for *animate* objects. *She* and *he* are English pronouns used for humans and often animals but are rarely used to refer to inanimate objects.[2] This results in an ambiguity when translating pronouns into English from languages that do not make this distinction. For example, in English, *she is in the kitchen* clearly refers to a person while *it is in the kitchen* refers to a non-person. In French, the word *elle* would be used in both situations, requiring an MT system to make a choice.

2.1.3 Formality

Social expectations dictate language usage. In many languages, this is explicitly lexicalized with different second-person pronouns and verb conjugations that distinguish intimate or familiar relationships from formal ones. Examples include the *tu/vous* distinction in French and *du/Sie* in German.

Over time, English has lost its formal register in pronouns (often called the T-V distinction) which other languages frequently employ. A common sentence "Where are you?" may have multiple

[2]A small exception occurs when inanimate objects are personified. A frequent example is boats, which are often referred to as *she* in English.

interpretations determined by the addressee, but subtle cues in preceding context may indicate the level of formality or familiarity of the speaker—a "sir", the domain, or profession mentioned can clarify this. When translating this sentence into French, the system must choose a register to produce either "Où êtes-vous?" or "Où es-tu?" There is often insufficient information to make the correct choice from just a single sentence.

2.2 Verb Phrase Ellipsis

Verb phrases can be dropped for emphasis, style, or brevity. The manner in which they are ellipsed will follow the rules of syntax of the specific language.

2.2.1 Isolated Auxiliaries

English auxiliaries ("do", "will", "would") can occur as standalone verbs by taking the place of a verb phrase. The question "Will you walk with me?" can be answered with a short "I will." Many target languages require translation of the original head of the verb phrase rather than the modal or auxiliary. Simply, "I will" must be translated as "I will walk" or rather "I walk" inflected in the future tense. We limit this work to the aforementioned auxiliaries as they rarely have direct translations.

2.2.2 Inflection of Verb-less Nouns

Extreme ellipsis may remove entire portions of a sentence and render it a *phrase*. English word order conveys grammatical role of nouns. When elements of the original sentence, such as the verb, are ellipsed, it may be impossible to infer the grammatical case of any remaining nouns which have no inflection. Translation into languages with case systems suffers. Voita et al. (2019) exemplifies using the phrase: "You call her your friend but have you been to her home? Her work?" To translate this phrase into Russian, it is necessary to know that "her work" has the same grammatical case as "her home" in the previous sentence.

3 Extraction Pipeline

Our pipeline functions by identifying up to four key tokens and ensuring each token matches a set of predefined criteria. The four components are: (1) The source (English) token defined as T_s, the target (non-English) token defined as T_t, the source token which conveys the contextual information required to resolve the ambiguity defined as C_s, and the target token aligned to C_s defined as C_t. These relationships are illustrated in Figure 1. Contextual information is defined by a contextual relationship, Q, which has an associated solver. The predefined criteria is a set of rules, R.

We can identify ambiguous sentences by:

1. For each source–target sentence pair, apply word alignment. Each aligned pair of words forms a potential T_s–T_t pair.
2. Ensure T_s meets all criteria R_{T_s}
3. Ensure T_t meets all criteria R_{T_t}
4. Apply a solver for the contextual relationship, Q to the English token T_s and its preceding context to identify C_s.
5. Ensure C_s meets all criteria R_{C_s}.
6. Identify the target token C_t via word alignment to C_s. If translation conveys semantic symmetry, this token *also* has a contextual relationship with T_t.
7. Ensure C_t meets all criteria R_{C_t}

Consider the ambiguity of pronoun resolution. Müller et al. (2018) first proposed a pipeline for extracting ambiguous translations of English "it" to German nominatives ("er", "es", and "sie"). We can explain their methodology[3] via the aforementioned definition. The following identifies all ambiguities where the English "it" is translated as "sie."

1. For each source-target sentence pair, apply word alignment. Each aligned pair of words forms a potential T_s–T_t pair.
2. Ensure T_s is the word "it"
3. Ensure T_t is the word "sie"
4. The contextual information to resolve the ambiguity is its antecedent—expressed via a coreference relationship. Apply a coreference resolver (Q) to identify C_s.
5. Ensure C_s is a noun (not another pronoun).
6. Identify C_t via word alignment.
7. Ensure C_t is a feminine, singular noun.

The same criteria could be enumerated for the masculine and neuter equivalents, appropriately changing gender and surface form checks.

To extract a specific phenomenon and language, a "rule" (R) must be written which specifies features that T_s, T_t, C_s, and C_t must have. These features can range from part-of-speech, lemma, gender, case, plurality or others. The manner in which

[3] Müller et al. (2018) performs an extra coreference check on the target side that we do not.

	English (T_e)			German (T_t)			Coref English (C_e)	Coref German (C_t)		
Rule	Form	POS	Case	Form	POS	Case	POS	POS	Gender	Number
NOM.FEM.SING	it	PNOUN	*	sie	PNOUN	Nom.	NOUN	NOUN	Fem.	Sing.
NOM.MASC.SING	it	PNOUN	*	er	PNOUN	Nom.	NOUN	NOUN	Masc.	Sing.
NOM.NEUT.SING	it	PNOUN	*	es	PNOUN	Nom.	NOUN	NOUN	Neut.	Sing.
ACC.FEM.SING	it	PNOUN	*	sie	PNOUN	Acc.	NOUN	NOUN	Fem.	Sing.
ACC.MASC.SING	it	PNOUN	*	ihn	PNOUN	Acc.	NOUN	NOUN	Masc.	Sing.
ACC.NEUT.SING	it	PNOUN	*	es	PNOUN	Acc.	NOUN	NOUN	Neut.	Sing.
DAT.FEM.SING	it	PNOUN	*	ihr	PNOUN	Dat.	NOUN	NOUN	Fem.	Sing.
DAT.MASC.SING	it	PNOUN	*	ihm	PNOUN	Dat.	NOUN	NOUN	Masc.	Sing.
DAT.NEUT.SING	it	PNOUN	*	ihm	PNOUN	Dat.	NOUN	NOUN	Neut.	Sing.
NOM.INFORM.SING	you	PNOUN	*	du	PNOUN	Nom.	-	-	-	-
NOM.FORM+PLUR	you	PNOUN	*	Sie	PNOUN	Nom.	-	-	-	-
NOM.INFORM.PLUR	you	PNOUN	*	ihr	PNOUN	Nom.	-	-	-	-
ACC.INFORM.SING	you	PNOUN	*	dich	PNOUN	Acc.	-	-	-	-
ACC.FORM+PLUR	you	PNOUN	*	Sie	PNOUN	Acc.	-	-	-	-
ACC.INFORM.PLUR	you	PNOUN	*	euch	PNOUN	Acc.	-	-	-	-
DAT.INFORM.SING	you	PNOUN	*	dir	PNOUN	Dat.	-	-	-	-
DAT.FORM+PLUR	you	PNOUN	*	ihnen	PNOUN	Dat.	-	-	-	-
DAT.INFORM.PLUR	you	PNOUN	*	euch	PNOUN	Dat.	-	-	-	-

Table 3: German criteria for all pronouns. We expand from Müller et al. (2018) to consider more cases (Accusative and Dative). English case is not used since the German annotations are more precise (English does not label Dative). PNOUN check in some cases is required to eliminate determiners (possessive adjectives instead of possessive pronouns)

	English (T_e)	French (T_t)
Rule	Lemma	Illegal Lemmas
DO.ELL	do	faire, aller
WOULD.ELL	would	faire, pouvoir
WILL.ELL	will	aller, faire

Table 4: French ellipsis Rules. English must have specified lemma. French alignment cannot have a lemma in the specified list.

these four components are identified creates the adaptability for each phenomena.

Gender Following previous works, we retrieve T_s and T_t based on surface form and word alignment. C_s is a noun discovered via coreference chain. If the coreference is a noun phrase, the head of the phrase is used. C_t is retrieved via word alignment. C_t must match the same morphological features present in T_t (e.g., gender and number).

Animacy As explained in Section 2.1.2, the animacy ambiguity that we consider occurs when translating from the gendered languages *into* English (whereas the gender ambiguity occurs when translating *out-of* English). To extract these examples, we use the same rules as GENDER, but we reverse the language direction for inference.

Formality The distinction of formality is the lack of a consistent or discrete C_s which informs the level of formality. Translating between English and a T-V language is always ambiguous with respect to the second person so we forgo using a contextual resolver Q to identify the appropriate context.

Auxiliary T_s is extracted from a pre-constructed list of auxiliaries—similar to those mentioned in Section 2.2.1. T_t, identified via word alignment, cannot occur in a pre-constructed list of forbidden translations. These translations are meant to prevent valid translations of auxiliaries, rather than the ambiguous ellipsed forms. For example, "to do" translated as a form of "faire" in French, is a direct translation, and is likely not representative of an ellipsed form. Contrarily, "to do" translated as a form of "savoir" in French is not a direct translation and is indicative of a previous occurrence of English "to know." C_t can be identified by finding the most recent occurrence of the same verb T_t, and C_s is retrieved from word alignment with C_t.

Inflection T_s and T_t can be of any form and any case. Any aligned noun pair (T_s and T_t) that occurs without an accompanying verb is ambiguous. C_t is identified as the most recent occurrence of *any noun* occurring in the same case as T_t. We assume the verb phrase surrounding C_t was ellipsed when generating T_t. We align C_t to find C_s.

We use FastCoref (Otmazgin et al., 2022) to perform English coreference resolution, simalign (Jalili Sabet et al., 2020) to perform cross-lingual

	DE	FR	RU	PL	PT	IT	ES
GENDER	147k	291k	113k	117k	127k	36k	96k
ANIMACY *	80k	145k	66k	39k	38k	20k	84k
FORMALITY	3.9M	5.7M	3.6M	1.7M	857k	833k	10.1M
AUXILIARY	4414	27.6k	39.1k	34.2k	30.2k	17.5k	29.6k
INFLECTION	-	-	2.6M	3.2M	-	-	-
# LINES	22.5M	41.9M	25.9M	77.2M	33.2M	35.2M	61.4M
% EXTRACTED	18%	14%	25%	6.6%	3.1%	2.5%	16.7%
%-COREFERENCE	0.7%	0.8%	0.6%	0.2%	0.5%	0.2%	0.2%

Table 5: OpenSubtitles2018 Extraction Statistics for each category. # LINES indicates the total number of lines in OpenSubtitles for the EN-XX language pair. % EXTRACTED indicates the percent of the dataset that was extracted. %-COREFERENCE indicates the classes that require a strict antecedent (GENDER and AUXILIARY). (*) ANIMACY was created by reversing a subset of the GENDER class so it is not used to calculate EXTRACTED because of the overlap.

word alignment, and SpaCy[4] to extract all other morphological features. We provide a larger list of our criteria in Appendix A.

3.1 Application to OpenSubtitles

We apply our extractor to the OpenSubtitles2018 dataset (Lison and Tiedemann, 2016) following previous work (Müller et al., 2018; Lopes et al., 2020). It comprises conversational dialog extracted from film and television subtitles. The conversational nature means plenty of context-based phenomenon occur. In Table 5, we present the total number of instances we extracted from Open Subtitles.

The fraction of the dataset that contains the phenomenon we target varies from language to language. This stems from the number of forms in each language, the number of genders, as well as translation standards. German, for instance, has very few AUXILIARY examples. We speculate this is due to German having similar auxiliary features as English so many examples were filtered out due to our "forbidden translation" criteria.

Some categories are extremely common. FORMALITY is invoked every time the second-person is used, which is frequent in conversational speech. INFLECTION also has high occurrences since there was relatively little filtering on the extracted examples. GENDER and AUXILIARY are *very rarely* extracted—less than 1% of the time in all languages. A 1% error rate is extreme when deploying at scale. Further, test sets, in nature, are small. If only 1% of the test set challenges contextual models, the results may be insignificant.

To form the dev, devtest, and test splits, we apply the following approach. For each label within a category, we ensure there are at least 100 examples. If there are fewer, we keep all examples for test. If there are more, we split the most recent years of OpenSubtitles into a 1:1:5 ratio for dev:devtest:test, limiting the test set's maximum size to 5000 examples per label. One label is roughly one surface form, but corresponds to one "rule" (a set of criteria R) or one row as shown in Table 3.

4 Quantitative Evaluation

Our goal is to show that our test sets can usefully discriminate between sentence-level and context-aware systems. An impediment to this goal is the lack of contextual machine translation models across languages for use in comparison and evaluation, and the difficulty in building them. Consequently, we turn to a commercial system, DeepL, which is alone among commercial providers in advertising contextual translation.[5] We translate with document-context by providing DeepL with context when translating, and compare to the same model translating without context at the sentence level. We show that a contextual system appropriately benefits from the additional context and gains significance performance on this test set.

Many works release their evaluation sets with the assumption of contrastive evaluation (Müller et al., 2018; Lopes et al., 2020; Voita et al., 2019), where the test is whether a model assigns a higher score to correct data than to linguistically-manipulated counterparts. This assumption ignores the fact that machine translation is a generative problem and should be evaluated as such. Recent work (Post and Junczys-Dowmunt, 2023) confronts this problem

[4] https://spacy.io/usage/models#languages

[5] https://www.deepl.com/docs-api/general/working-with-context

		Generative Accuracy (%)							COMET						
		DE	ES	FR	IT	PL	PT	RU	DE	ES	FR	IT	PL	PT	RU
GENDER	sent.	48.1	34.6	40.2	51.1	32.8	44.3	35.9	0.23	0.50	0.33	0.43	0.51	0.52	0.36
	doc.	**73.3**	**47.4**	**59.0**	**68.3**	**50.2**	**64.3**	**51.8**	**0.31**	**0.52**	**0.43**	**0.48**	**0.54**	**0.57**	**0.42**
		+25.2	+12.8	+18.8	+17.2	+17.4	+20.0	+15.9	+0.08	+0.02	+0.09	+0.05	+0.03	+0.05	+0.06
ANIMACY	sent.	61.0	84.4	68.0	81.4	57.6	64.1	55.4	0.27	0.53	0.40	0.42	0.25	0.43	0.19
	doc.	**74.1**	**87.8**	**75.2**	**86.1**	**70.5**	**79.5**	**71.6**	**0.38**	**0.58**	**0.49**	**0.46**	**0.31**	**0.55**	**0.34**
		+13.1	+3.4	+7.2	+4.7	+12.9	+15.4	+16.2	+0.11	+0.05	+0.09	+0.04	+0.06	+0.12	+0.15
FORMALITY	sent.	44.0	31.7	40.6	38.9	25.3	40.1	55.4	**0.32**	0.54	0.45	0.47	**0.51**	**0.59**	0.57
	doc.	**53.6**	**35.9**	**51.5**	**46.1**	**31.6**	**47.2**	**62.5**	**0.32**	**0.55**	**0.48**	**0.48**	**0.51**	**0.59**	**0.58**
		+9.6	+4.2	+10.9	+7.2	+6.3	+7.1	+7.1	+0.0	+0.01	+0.03	+0.01	+0.0	+0.0	+ 0.01
AUXILIARY	sent.	7.8	3.3	1.3	4.0	8.2	9.2	5.7	-0.27	-0.06	-0.34	-0.02	0.10	0.03	-0.09
	doc.	**40.0**	**52.0**	**32.2**	**40.7**	**49.9**	**53.8**	**49.0**	**0.04**	**0.54**	**0.20**	**0.38**	**0.53**	**0.60**	**0.49**
		+32.2	+48.7	+30.9	+36.7	+41.7	+44.6	+43.3	+0.31	+0.60	+0.54	+0.40	+0.43	+0.57	+0.58
INFLECTION	sent.	-	-	-	-	41.3	-	34.6	-	-	-	-	0.57	-	0.47
	doc.	-	-	-	-	**53.2**	-	**48.3**	-	-	-	-	**0.68**	-	**0.56**
		-	-	-	-	+11.9	-	+13.7	-	-	-	-	+0.11	-	+0.09

Table 6: Generative evaluation percent accuracy scores (left section) evaluation ability to produce expected form; COMET scores (right section) evaluate the translation quality of this model; *sent.* denotes that no additional context was given while *doc.* was given five consecutive sentences for context. All translations made using DeepL commercial API. ANIMACY is *into* English. All others are *out of* English

and proposes generative evaluation as an alternative, showing a wide gap between contextual and sentence-level systems that is only observed under generative evaluation. Translations are counted as correct if the *expected surface form* is present anywhere in the model's output—matching the entire word and not simply a substring. We follow this approach in our evaluation.

We validate our data by showing it (1) adequately addresses context-based phenomena and (2) is sufficiently challenging. We demonstrate the former by showing that a context-aware translation model consistently outperforms a context-less equivalent. We see the latter is true as the contextual model does not solve the problem. There is still significant context-aware work to be done.

4.1 Accuracy

We begin by translating sentences both with and without context, using at most five sentences of context. To limit API calls, we run a subsample of our produced evaluation sets. We limit each category (GENDER, ANIMACY, FORMALITY, AUXILIARY, and INFLECTION) to approximately 10k total examples, divided evenly amongst the categories labels. To extract the final sentence for scoring purposes, we apply segmentation using the ERSATZ segmenter (Wicks and Post, 2022).

The results in Table 6 clearly show that the DeepL model with additional context *far outperforms* its sentence-level equivalent.[6] Many of these evaluation examples have specific preceding context that needs to be used in order to correctly translate the ambiguity. FORMALITY is a slight exception. There is little to no guarantee that explicit cues are given to convey the nature of the relationship between the speaker and addressee, yet preceding context still benefits an average of 9 percentage points across all languages. AUXILIARY is a task of translating verbs. A random guess would equate to sampling from the distribution of verbs in a language–which results in low success rates. Translating AUXILIARY with context increases from nearly never correct to a roughly 50% accuracy rate. Translating ANIMACY has higher sentence-level baselines than some of the other categories. We attribute this to other semantic cues towards ANIMACY which are less arbitrary than something such as GENDER assignment. For instance, if a noun *talks*, it is likely animate, while a noun that *is thrown* is likely inanimate. Similarly, INFLECTION may have some sentence-internal cues. Certain nouns may have a majority class, or preceding prepositions ((“with”, “for”, “in”, etc.) may indicate case. This is similar to the intrasentential coreference found with pronouns, which makes some occurrences easier than oth-

[6] Ideally we would make the same comparison between document- and sentence-level translation with other commercial systems, but there is no way to prevent them from applying sentence-level segmentation to the document-context string.

ers. Nonetheless, additional context aids the model. In every category, the context-aware model shows consistent gains over its context-less variant.

4.2 Automatic metric

We also present COMET scores (Rei et al., 2020) in Table 6. Across all categories and language pairs, COMET shows improvement when the system leverages additional context. The consistent improvement in COMET reinforces the trends we see with the generative evaluation metric. The one exception is the FORMALITY class which has minimal differences between the sentence-level and contextual inputs. COMET rewards synonyms and we suspect formal and informal surface forms have more similar encodings in COMET models than these other grammatical forms. A surface-based metric would better capture the gains that can be seen from the accuracy scores, which is indeed what we find (Table 18 in Appendix A).

5 Qualitative Evaluation

Our extraction pipeline relies on handbuilt rules applied to the outputs of automatic tools. As a result, the process is noisy and may be susceptible to errors. The previous section showed that a contextual system does better on our test sets than its sentence-based counterpart, and there is no reason we can think of to suspect that errors would systematically benefit the contextual system. However, in the interest of completeness, we took a more qualitative look at the data. This includes a systematic manual review (§ 5.1), direct comparison with prior work (§ 5.2), and an error analysis (§ 5.3).

5.1 Manual review

Previous work in automatic test set production has not typically included a manual analysis of rule quality. To build confidence in these automatic extraction methodologies, we sampled 100 random test examples from the extracted English–French GENDER set and manually reviewed and annotated them for errors. We find that 92 of the extracted examples are correct. Three more were questionably incorrect—with correct translations and alignments—yet had atypical coreference resolutions that were difficult for our human reviewer to understand. Of the remaining five, two had a non-referential pronoun. One such example "What is it?" was used in the sense of "What's wrong?" rather than "What is that?" In the former, "it" has no valid antecedent, yet it was extracted.

We present the remaining three errors in Table 7, where they demonstrate where errors arise at each step in the pipeline. The Coreference Error is a clear mistake. "They don't want us to know what they're working on" refers to the people being talked to, and not "these guys"—who instead seemed to be criminals who broke into a company. The Alignment Error is an unfortunate combination of a bad alignment and inconsistent translation. "the discipline" is aligned to the word "espionnage." "discipline" in French is a feminine noun, while "espionnage" is masculine. The French "il" is masculine, and thus has "espionnage" as an antecedent despite the English having "the discipline." This coincidental error caused this example to still be extracted. Lastly, one of these examples seemed to have a typo in the English transcript. The word "signatures" seemed to be incorrect. We suspect the correct transcription word was "serial killers." Given the inconsistent context on the English side, we suspect the neural coreference model had difficulties resolving this.

5.2 Comparison with prior work

Since our extraction framework is largely based on that of Müller et al. (2018), we expect to have a similar quality of extracted rules (or better, since the underlying annotations tools have improved). We thus undertake a comparison to the data that they released. When applying our pipeline to the German–English OpenSubtitles data, we extract 147,211 sentences that have ambiguous pronoun usage. Müller did not report their raw extraction numbers, but their release includes 12,000 examples, balanced across gender (but not distance). We therefore focus our analysis on this subset.

Since their pipeline contained a target-side coreference check that we do not have, one might think their pipeline would be a stricter selection process, but we find the opposite to be true. Our pipeline's selection overlaps with only half of ContraPro (6,003 sentences), rejecting the other half (5,997 sentences). An analysis of this rejected portion of ContraPro turns up some explanations. ContraPro extracts three categories of German pronouns corresponding to neuter, masculine, and feminine genders. For *er* and *sie*, we rejected roughly 25% of the ContraPro examples; however, we rejected over 75% of the neuter examples from ContraPro. Upon review, we found a substantial number of

Error Type	English	French
Coreference	We got any ideas what these ***guys*** were after?	Une idée de ce que voulaient ces ***gars***?
	No, CEO is on his way down to talk to us now.	Non, le PDG arrive pour nous le dire.
	So far, everyone we've talked to hasn't really given us much.	Tous ceux à qui on a parlés ne nous ont rien appris.
	Makes sense.	C'est logique.
	They don't want us to know what they're working on here.	***Ils*** ne veulent pas qu'on sache ce qu'ils font.
Alignment	As you know the ***discipline*** of media espionage is a new one.	Comme vous le savez, l'***espionnage*** médiatique est une nouvelle discipline.
	Oh yes, ***it*** is everywhere.	***Il*** est partout.
Translation	You know more about ***signatures*** than most of ***them*** put together.	Vous en savez plur sur ces ***tueurs*** (`en: killers`) qu'***eux*** tous réunis

Table 7: In a sample of 100 extracted items, 8 errors were found. This table shows 3 of these errors made by the extraction pipeline on the French Gender set. The ***indicated*** words show the pronouns in French and English, as well as their antecedents. Some examples fit into multiple categories, but these show the most evident error type. `en:` indicates the English translation of French word.

non-referential instances. These examples include sentences such as "It was your duty.", "It would have been all right if it wasn't for you." and "It was one of those California Spanish houses" that all have either a non-specified referent or have a passive construction. The inclusion of these examples points to inaccurate coference chains, likely explained by their use of older corefence tools.

Our extraction employs strict criteria to find the head of a span during coreference and alignment. The head is used for the gender, person, and number checks included in the definition of R (§ 3). From our understanding of Müller's work, they did not include this check. Mistakes are inherent to any automatic process, and likely persist in our dataset as well. Our analysis here lends some confidence to the belief that tighter selection criteria and improved underlying tools result in better data.

5.3 Model analysis

Absent sufficient information, the translation of ambiguous words will regress to their proportions in the training data. For pronouns, this would be the neuter or masculine class; for auxiliaries, the direct translation (the "Illegal Lemma" in R).

We examine the English–German model outputs. Our evaluation sets have balanced counts across genders, so a correct model would produce a neuter pronoun roughly one-third of the time. Instead, this sentence-level model produces either "es" or "ihm" (the German neuter pronouns) closer to two-thirds of the time. This contextual model has better performance producing the neuter pronouns about 40% of the time. This problem is well-known, but other issues are not as well documented.

The auxiliary category had the worst scores, both in terms of how low the sentence-level model was performing as well as the absolute increase from adding context. The cause of these scores becomes obvious as we examine the model outputs. To generate the rules for the AUXILIARY class, we enumerated illegal lemmas that represent the most common direct translations of English modals as described in Section 3. Ideally, a model would never generate these verbs for our evaluation set unless part of a larger verb phrase construction. We find the sentence-level model generates a translation that contains a form of one of these lemmas approximately two-thirds of the time. Conversely, the contextual model generates these closer to one-third of the time.

6 Analysis of WMT test sets

As previously earlier, this pipeline is easily applied to new data and test sets. We demonstrate this by applying it to the 2019–2022 WMT newswire test sets (Barrault et al., 2019, 2020; Akhbardeh et al., 2021; Kocmi et al., 2022). In so doing, we find

	DE	RU	PL
GENDER	135	64	13
FORMALITY	540	416	4
AUXILIARY	1	0	0
INFLECTION	-	14	1
WMT # LINES	6454	7038	1000

Table 8: Counts on the number of extracted examples from WMT 2019-2022 (when available) test sets.

phenomena in a similar proportion of sentences to OpenSubtitles, but with a different distribution; there is a higher rate of GENDER but smaller of FORMALITY and AUXILIARY. In Table 8, we present the total number of examples discovered in WMT 2019-2022 in `en-de`, `en-ru`, and `en-pl` (when available). The newswire text hardly ever contains the AUXILIARY type of ambiguity. Formality comprises the bulk of the examples, and upon further inspection, we find a severe bias towards the formal register, with a 1 to 7 ratio of informal to formal—likely due to the characteristics of the domain. Further, we suspect the sparseness in contextual ambiguities is important to consider when evaluating these systems.

7 Related Work

Work in contextual machine translation can be divided into three categories: (1) the publication of resources, similar to this work; (2) alterations on the training paradigm via architecture or data input; (3) evaluation metrics.

This work largely follows the path set forth by those who have previously published resources on the detection of gender, pronouns, and formality (Guillou and Hardmeier, 2016; Müller et al., 2018; Bawden et al., 2018; Voita et al., 2019; Lopes et al., 2020). (Currey et al., 2022) produces a gender-based evaluation dataset using human annotators, but covers the complement of this work: gender assigned to humans rather than inanimate objects. In addition to the manual pipelines, recent work has been done to promote the automatic detection of these phenomena. Nadejde et al. (2022) implements a cross-lingual mutual information metric that tags words as needing additional context. The tags were found to often overlap with the variety discussed in this work. Fernandes et al. (2023) also use a mutual-information based score to select data that is then used to derive a similar rule-based extraction approach, but do not release evaluation sets.

A substantial amount of work has been done to allow traditional neural models to handle additional input. Some approaches involve more complex architectures or modifications to training paradigms incorporate longer sequences (Miculicich et al., 2018; Bao et al., 2021), but Sun et al. (2022) showed that unaltered Transformers can handle longer sequences. Other work has focused on leveraging and cleaning the available data, since large-scale document bitext is lacking (Junczys-Dowmunt, 2019; Post and Junczys-Dowmunt, 2023).

Lastly, many have realized that BLEU, COMET, or other sentence-level metrics will not address the distinction in document-level performance. Vernikos et al. (2022) proposed a new method for adjusting current methods to adjust for document-level inputs. Jiang et al. (2022) proposed BlonDe, an entirely novel metric for document-level evaluation. We hope this work complements these works and serves to further the field in its aspirations towards true context-aware translation.

8 Summary

Machine translation systems face a performance ceiling that can't be overcome so long as they continue to operate at the sentence level. A major obstacle to that transition is the unavailability of test sets in many languages and for many contextual phenomena. The goal of this work has been to help address that problem. The extraction pipeline proposed in this paper can be used to identify and generate new test sets which contain linguistic phenomena that can only be consistently translated by contextual systems. The application of our pipeline to the OpenSubtitles dataset in seven languages provides a new set of evaluation sets including a wider set of languages and phenomena than were available before. Further, we hope that the extensibility of our pipeline to new phenomena and languages allows for others to build upon this work to expand resources and coverage. The CTXPRO datasets and extraction pipeline are available as open source from `https://github.com/rewicks/ctxpro`.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa,

Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Guangsheng Bao, Yue Zhang, Zhiyang Teng, Boxing Chen, and Weihua Luo. 2021. G-transformer for document-level machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3442–3455, Online. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Anna Currey, Maria Nadejde, Raghavendra Reddy Pappagari, Mia Mayer, Stanislas Lauly, Xing Niu, Benjamin Hsu, and Georgiana Dinu. 2022. MT-GenEval: A counterfactual and contextual dataset for evaluating gender accuracy in machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4287–4299, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Patrick Fernandes, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig. 2023. When does translation require context? a data-driven, multilingual exploration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 606–626, Toronto, Canada. Association for Computational Linguistics.

Liane Guillou and Christian Hardmeier. 2016. PROTEST: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 636–643, Portorož, Slovenia. European Language Resources Association (ELRA).

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at wmt 2019: Towards large-scale document-level neural machine translation.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from

movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

Maria Nadejde, Anna Currey, Benjamin Hsu, Xing Niu, Marcello Federico, and Georgiana Dinu. 2022. CoCoA-MT: A dataset and benchmark for contrastive controlled MT with application to formality. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 616–632, Seattle, United States. Association for Computational Linguistics.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2022. F-coref: Fast, accurate and easy to use coreference resolution.

Matt Post and Marcin Junczys-Dowmunt. 2023. Escaping the sentence-level paradigm in machine translation.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. Rethinking document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3537–3548, Dublin, Ireland. Association for Computational Linguistics.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Rachel Wicks and Matt Post. 2022. Does sentence segmentation matter for machine translation? In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 843–854, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

A Additional Materials

	English (T_e)	German (T_t)
Rule	Lemma	Illegal Lemmas
DO.ELL	do	machen, tun, haben, können
WOULD.ELL	would	machen, tun, haben
WILL.ELL	will	machen, tun, haben, werden

Table 9: German auxiliary rules. English must have specified lemma. German alignment cannot have a lemma in the specified list.

	English (T_e)	Polish (T_t)
Rule	Lemma	Illegal Lemmas
DO.ELL	do	robić
WOULD.ELL	would	robić, by być, być, by, móc
WILL.ELL	will	robić, by być, być, by, móc, iść

Table 10: Polish auxiliary rules. English must have specified lemma. Polish alignment cannot have a lemma in the specified list.

	English (T_e)	Russian (T_t)
Rule	Lemma	Illegal Lemmas
DO.ELL	do	Делать
WOULD.ELL	would	Делать
WILL.ELL	will	Делать

Table 11: Russian auxiliary rules. English must have specified lemma. Russian alignment cannot have a lemma in the specified list.

	English (T_e)	Portugese (T_t)
Rule	Lemma	Illegal Lemmas
DO.ELL	do	fazer
WOULD.ELL	would	fazer, poder
WILL.ELL	will	fazer, ir

Table 12: Portuguese auxiliary rules. English must have specified lemma. Portuguese alignment cannot have a lemma in the specified list.

	English (T_e)	Italian (T_t)
Rule	Lemma	Illegal Lemmas
DO.ELL	do	fare
WOULD.ELL	would	fare, potere, volere
WILL.ELL	will	fare, andare

Table 13: Italian auxiliary rules. English must have specified lemma. Italian alignment cannot have a lemma in the specified list.

	English (T_e)			Spanish (T_t)			Coref English (C_e)	Coref Spanish (C_t)		
Rule	Form	POS	Case	Form	POS	Case	POS	POS	Gender	Number
NOM.FEM.SING	it	PNOUN	Nom.	ella	PNOUN	*	NOUN	NOUN	Fem.	Sing.
NOM.MASC.SING	it	PNOUN	Nom.	él	PNOUN	*	NOUN	NOUN	Masc.	Sing.
NOM.FEM.PLUR	it	PNOUN	Nom.	ellas	PNOUN	*	NOUN	NOUN	Fem.	Plur.
NOM.MASC.PLUR	it	PNOUN	Nom.	ellos	PNOUN	*	NOUN	NOUN	Masc.	Plur.
ACC.MASC.SING	it	PNOUN	Acc.	lo	PNOUN	*	NOUN	NOUN	Masc.	Sing.
ACC.FEM.SING	it	PNOUN	Acc.	la	PNOUN	*	NOUN	NOUN	Fem.	Sing.
ACC.MASC.PLUR	them	PNOUN	Acc.	los	PNOUN	*	NOUN	NOUN	Masc.	Sing.
ACC.FEM.PLUR	them	PNOUN	Acc.	las	PNOUN	*	NOUN	NOUN	Fem.	Sing.
DISJ.MASC.SING	it	PNOUN	-Nom.	él	PNOUN	*	NOUN	NOUN	Masc.	Sing.
DISJ.MASC.SING.ALT	it	PNOUN	-Nom	ello	PNOUN	*	NOUN	NOUN	Masc.	Sing.
DISJ.FEM.SING	it	PNOUN	-Nom	ella	PNOUN	*	NOUN	NOUN	Fem.	Sing.
DISJ.MASC.PLUR	them	PNOUN	-Nom	ellos	PNOUN	*	NOUN	NOUN	Masc.	Plur.
DISJ.FEM.PLUR	them	PNOUN	-Nom	ellas	PNOUN	*	NOUN	NOUN	Fem.	Plur.
NOM.INFORM.SING	you	PNOUN	Nom.	tú	PNOUN	*	-	-	-	-
NOM.FORM.SING	you	PNOUN	Nom.	usted	PNOUN	*	-	-	-	-
NOM.FORM.PLUR	you	PNOUN	Nom.	ustedes	PNOUN	*	-	-	-	-
NOM.INFORM.PLUR.MASC	you	PNOUN	Nom.	vosotros	PNOUN	*	-	-	-	-
NOM.INFORM.PLUR.FEM	you	PNOUN	Nom.	vosotras	PNOUN	*	-	-	-	-
ACC.INFORM.SING	you	PNOUN	Acc.	te	PNOUN	*	-	-	-	-
ACC.FORM.SING.MASC	you	PNOUN	Acc.	lo	PNOUN	*	-	-	-	-
ACC.FORM.SING.FEM	you	PNOUN	Acc.	la	PNOUN	*	-	-	-	-
ACC.FORM.PLUR.MASC	you	PNOUN	Acc.	los	PNOUN	*	-	-	-	-
ACC.FORM.PLUR.FEM	you	PNOUN	Acc.	las	PNOUN	*	-	-	-	-
ACC.INFORM.PLUR	you	PNOUN	Acc.	os	PNOUN	*	-	-	-	-
DISJ.INFORM.SING	you	PNOUN	-Nom.	ti	PNOUN	*	-	-	-	-
DISJ.INFORM.SING.ALT	you	PNOUN	-Nom.	contigo	PNOUN	*	-	-	-	-
DISJ.FORM.SING	you	PNOUN	-Nom.	usted	PNOUN	*	-	-	-	-
DISJ.INFORM.PLUR.MASC	you	PNOUN	-Nom.	vosotros	PNOUN	*	-	-	-	-
DISJ.INFORM.PLUR.FEM	you	PNOUN	-Nom.	vosotras	PNOUN	*	-	-	-	-
DISJ.FORM.PLUR	you	PNOUN	-Nom.	ustedes	PNOUN	*	-	-	-	-

Table 14: Spanish Pronoun Rules

	English (T_e)			French (T_t)			Coref English (C_e)	Coref French (C_t)		
Rule	Form	POS	Case	Form	POS	Case	POS	POS	Gender	Number
NOM.FEM.SING	it	PNOUN	Nom.	elle	PNOUN	*	NOUN	NOUN	Fem.	Sing.
NOM.MASC.SING	it	PNOUN	Nom.	il	PNOUN	*	NOUN	NOUN	Masc.	Sing.
NOM.FEM.PLUR	they	PNOUN	Nom.	elles	PNOUN	*	NOUN	NOUN	Fem.	Plur.
NOM.MASC.PLUR	they	PNOUN	Nom.	ils	PNOUN	*	NOUN	NOUN	Masc.	Plur.
ACC.MASC.SING	it	PNOUN	Acc.	le	PNOUN	*	NOUN	NOUN	Masc.	Sing.
ACC.FEM.SING	it	PNOUN	Acc.	la	PNOUN	*	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.1S	mine	PNOUN	*	mienne	PNOUN	*	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.1P	ours	PNOUN	*	la nôtre	PNOUN	*	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.2S	yours	PNOUN	*	tienne	PNOUN	*	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.2P	yours	PNOUN	*	la vôtre	PNOUN	*	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.3SM	his	PNOUN	*	sienne	PNOUN	*	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.3SF	hers	PNOUN	*	sienne	PNOUN	*	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.3N	its	PNOUN	*	sienne	PNOUN	*	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.3P	theirs	PNOUN	*	la leur	PNOUN	*	NOUN	NOUN	Fem.	Sing.
NOM.INFORM.SING	you	PNOUN	Nom.	tu	PNOUN	*	-	-	-	-
NOM.FORM+PLUR	you	PNOUN	Nom.	vous	PNOUN	*	-	-	-	-
ACC.INFORM.SING	you	PNOUN	Acc.	te	PNOUN	*	-	-	-	-
ACC.INFORM.SING.LIAS	you	PNOUN	Acc.	t’	PNOUN	*	-	-	-	-
ACC.FORM+PLUR	you	PNOUN	Acc.	vous	PNOUN	*	-	-	-	-
DISJ.INFORM.SING	you	PNOUN	-Nom	toi	PNOUN	*	-	-	-	-

Table 15: A sampling of French pronoun rules (abridged). Some forms left off for space.

	English (T_e)			Italian (T_t)			Coref English (C_e)	Coref Italian (C_t)		
Rule	Form	POS	Case	Form	POS	Case	POS	POS	Gender	Number
NOM.MASC.SING	it	PNOUN	Nom.	lui	PNOUN	*	NOUN	NOUN	Masc.	Sing.
NOM.FEM.SING	it	PNOUN	Nom.	lei	PNOUN	*	NOUN	NOUN	Fem.	Sing.
ACC.MASC.SING	it	PNOUN	Acc.	lo	PNOUN	*	NOUN	NOUN	Masc.	Sing.
ACC.FEM.SING	it	PNOUN	Acc.	la	PNOUN	*	NOUN	NOUN	Fem.	Sing.
ACC.MASC.PLUR	them	PNOUN	Acc.	li	PNOUN	*	NOUN	NOUN	Masc.	Plur.
ACC.FEM.PLUR	them	PNOUN	Acc.	le	PNOUN	*	NOUN	NOUN	Fem.	Plur.
DAT.MASC.SING	it	PNOUN	Acc.	gli	PNOUN	*	NOUN	NOUN	Masc.	Sing.
DAT.FEM.SING	it	PNOUN	Acc.	le	PNOUN	*	NOUN	NOUN	Fem.	Sing.
DISJ.MASC.SING	it	PNOUN	-Nom	lui	PNOUN	*	NOUN	NOUN	Masc.	Sing.
DISJ.FEM.SING	it	PNOUN	-Nom	lei	PNOUN	*	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.1S	mine	PNOUN	*	mia	PNOUN	*	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.2S	yours	PNOUN	*	tua	PNOUN	*	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.3M	his	PNOUN	*	sua	PNOUN	*	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.3F	hers	PNOUN	*	sua	PNOUN	*	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.3N	its	PNOUN	*	sua	PNOUN	*	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.2P	yours	PNOUN	*	vostra	PNOUN	*	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.3P	theirs	PNOUN	*	loro	PNOUN	*	NOUN	NOUN	Fem.	Sing.
NOM.INFORM.SING	you	PNOUN	*	tu	PNOUN	*	-	-	-	-
NOM.FORM.SING	you	PNOUN	*	lei	PNOUN	*	-	-	-	-
NOM.INFORM.PLUR	you	PNOUN	*	voi	PNOUN	*	-	-	-	-

Table 16: A sampling of Italian Pronoun Rules. We do not consider the conflated Italian pronouns which combine accusatives and datives which co-occur. English case is used as it is a better model. Accusative is used for dative since the SpaCy models conflate the two in English.

	English (T_e)			Polish (T_t)			Coref English (C_e)	Coref Polish (C_t)		
Rule	Form	POS	Case	Form	POS	Case	POS	POS	Gender	Number
NOM.NEUT.SING	it	PNOUN	*	ono	PNOUN	Nom.	NOUN	NOUN	Neut.	Sing.
NOM.MASC.SING	it	PNOUN	*	on	PNOUN	Nom.	NOUN	NOUN	Masc.	Sing.
NOM.FEM.SING	it	PNOUN	*	ona	PNOUN	Nom.	NOUN	NOUN	Fem.	Sing.
ACC.NEUT.SING	it	PNOUN	*	je	PNOUN	Acc.	NOUN	NOUN	Neut.	Sing.
ACC.NEUT.SING.ALT1	it	PNOUN	*	nie	PNOUN	Acc.	NOUN	NOUN	Neut.	Sing.
ACC.MASC.SING	it	PNOUN	*	je	PNOUN	Acc.	NOUN	NOUN	Masc.	Sing.
ACC.MASC.SING.ALT	it	PNOUN	*	niego	PNOUN	Acc.	NOUN	NOUN	Masc.	Sing.
ACC.FEM.SING	it	PNOUN	*	ją	PNOUN	Acc.	NOUN	NOUN	Fem.	Sing.
GEN.NEUT.SING	it	PNOUN	*	jego	PNOUN	Gen.	NOUN	NOUN	Neut.	Sing.
GEN.NEUT.SING.ALT1	it	PNOUN	*	niego	PNOUN	Gen.	NOUN	NOUN	Neut.	Sing.
GEN.NEUT.SING.ALT2	it	PNOUN	*	go	PNOUN	Gen.	NOUN	NOUN	Neut.	Sing.
GEN.MASC.SING	it	PNOUN	*	je	PNOUN	Gen.	NOUN	NOUN	Masc.	Sing.
GEN.MASC.SING.ALT1	it	PNOUN	*	niego	PNOUN	Gen.	NOUN	NOUN	Masc.	Sing.
GEN.FEM.SING	it	PNOUN	*	jej	PNOUN	Gen.	NOUN	NOUN	Fem.	Sing.
GEN.FEM.SING.ALT1	it	PNOUN	*	niej	PNOUN	Gen.	NOUN	NOUN	Fem.	Sing.
LOC.NEUT.SING	it	PNOUN	*	nim	PNOUN	Loc.	NOUN	NOUN	Neut.	Sing.
LOC.MASC.SING	it	PNOUN	*	nim	PNOUN	Loc.	NOUN	NOUN	Masc.	Sing.
LOC.FEM.SING	it	PNOUN	*	niej	PNOUN	Loc.	NOUN	NOUN	Fem.	Sing.
DAT.NEUT.SING	it	PNOUN	*	jemu	PNOUN	Dat.	NOUN	NOUN	Neut.	Sing.
INS.NEUT.SING	it	PNOUN	*	nim	PNOUN	Ins.	NOUN	NOUN	Neut.	Sing.
NOM.INFORM.SING	you	PNOUN	*	ty	PNOUN	Nom.	-	-	-	-
ACC.INFORM.SING	you	PNOUN	*	ciebie	PNOUN	Acc.	-	-	-	-
NOM.FORM.SING.FEM	you	PNOUN	*	pani	PNOUN	Nom.	-	-	-	-
ACC.FORM.SING.FEM	you	PNOUN	*	panią	PNOUN	Acc.	-	-	-	-

Table 17: A sampling of Polish Pronoun Rules. Some left off for space.

		DE	ES	FR	IT	PL	PT	RU
GENDER	sent.	29.0	35.4	32.6	28.7	23.8	27.8	24.7
	doc.	**33.8**	**38.7**	**37.2**	**32.7**	**27.1**	**31.3**	**27.6**
		+4.8	+4.6	+2.9	+3.3	+3.5	+4.0	+3.3
ANIMACY	sent.	33.3	44.3	37.5	35.1	29.8	40.5	32.1
	doc.	**37.7**	**48.3**	**40.6**	**37.6**	**32.3**	**44.4**	**36.0**
		+4.4	+4.0	+3.1	+2.5	+2.5	+3.9	+3.9
FORMALITY	sent.	26.4	32.0	28.4	21.7	36.1	29.2	34.3
	doc.	**28.4**	**35.6**	**30.2**	**23.4**	**37.1**	**31.3**	**36.1**
		+2.0	+3.6	+1.8	+1.7	+1.0	+2.1	+1.8
AUXILIARY	sent.	17.7	17.3	14.9	17.8	15.3	15.8	19.9
	doc.	**30.1**	**33.4**	**33.6**	**34.7**	**33.1**	**32.5**	**42.2**
		+12.4	+16.1	+18.7	+16.9	+17.8	+16.7	+22.3
INFLECTION	sent.	-	-	-	-	27.3	-	27.7
	doc.	-	-	-	-	**30.7**	-	**29.9**
		-	-	-	-	+2.2	-	+3.4

Table 18: BLEU scores to evaluate the translation quality of this model. Higher is better. *sent.* denotes that no additional context was given while *doc.* was given five consecutive sentences. All translations produced by DeepL commercial API.

Machine Translation with Large Language Models: Prompting, Few-shot Learning, and Fine-tuning with QLoRA

Xuan Zhang
Johns Hopkins University
xuanzhang@jhu.edu

Navid Rajabi
George Mason University
nrajabi@gmu.edu

Kevin Duh
Johns Hopkins University
kevinduh@cs.jhu.edu

Philipp Koehn
Johns Hopkins University
phi@jhu.edu

Abstract

While large language models have made remarkable advancements in natural language generation, their potential in machine translation, especially when fine-tuned, remains underexplored. In our study, we conduct comprehensive experiments, evaluating 15 publicly available language models on machine translation tasks. We compare the performance across three methodologies: zero-shot prompting, few-shot learning, and fine-tuning. Central to our approach is the use of QLoRA, an efficient fine-tuning method. On French-English, QLoRA fine-tuning outperforms both few-shot learning and models trained from scratch. This superiority is highlighted in both sentence-level and document-level translations, with a significant BLEU score improvement of 28.93 over the prompting method. Impressively, with QLoRA, the enhanced performance is achieved by fine-tuning a mere 0.77% of the model's parameters.

1 Introduction

The rapid advancement of large language models (LLMs) is reshaping the field of natural language processing (NLP), marking a potential paradigm shift in future development (Zhao et al., 2023). Instead of crafting dedicated task-specific systems, a growing interest has been focusing on quickly adapting LLMs to specific tasks simply through prompting (Liu et al., 2023; Sanh et al., 2022). So far, studies have shown that prompting LLMs can match or even rival the performance of specialized systems on numerous NLP tasks (Radford et al.).

Among all the NLP tasks, the application of LLMs to machine translation (MT) is understudied. The optimal way to harness LLMs for MT remains an open question. While encoder-decoder-based LLMs (Xue et al., 2021; Liu et al., 2020; Costa-jussà et al., 2022) are inherently designed for the sequence-to-sequence demands of MT, the approach for leveraging decoder-only models is less straightforward.

Although there are initial attempts in this direction (Sia and Duh, 2022; Hendy et al., 2023; Moslem et al., 2023; Zhu et al., 2023), these studies mainly concentrate on prompting and few-shot learning, not exploiting the availability of bitext. Additionally, most work focus on exceptionally large LLMs like GPT3 (Brown et al., 2020) with its staggering 175 billion parameters, which are beyond the reach of non-commercial research groups for local training. This poses a significant hurdle for institutions with constrained computational resources, rendering the findings less applicable and relevant to many researchers.

In this paper, we aim to investigate the performance of LLMs on MT tasks, with a particular focus on decoder-based LLMs, a category less charted for MT applications. Our research focuses on a range of publicly available medium-sized LLMs. This includes models pretrained on English-centric datasets, such as GPT-Neo (Black et al., 2021), OPT (Zhang et al., 2022), LLaMA2 (Touvron et al., 2023), as well as those on multilingual datasets such as XGLM (Lin et al., 2021) and BLOOMZ (Muennighoff et al., 2022). We evaluate various versions of these models, with their parameter sizes spanning from 1.3 billion to 13 billion, totaling 15 models.

In our experiments, we explore zero-shot prompting, few-shot learning, and fine-tuning, where our emphasis on fine-tuning fills the gap in previous studies. For the fine-tuning process, we employ the QLoRA method (Dettmers et al., 2023), which enhances efficiency and minimizes memory usage by quantizing the model to 4-bit precision and limiting the number of trainable parameters. To the best of our knowledge, this is the first instance of QLoRA being applied to fine-tuning LLMs for MT tasks.

We also evaluate the performance of LLMs in document-level translation. Standard sequence-to-sequence MT models focus on translating one sentence at a time, overlooking discourse phenom-

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 468–481
December 6–7, 2023.

ena and the broader context. Existing methods for document-level translation often pivot toward architectural modifications (Tu et al., 2018; Tan et al., 2019; Xu et al., 2021), leading to specialized models that need unique designs. Our objective is to evaluate the capability of LLMs in preserving long-term contextual coherence and to explore their potential in facilitating the development of a robust document-level translation system.

We demonstrate the effectiveness of fine-tuning on a French-English dataset – this language pair is selected due to its accessibility for LLMs, positioning it as an ideal starting point for research in this domain. Our experimental results, complemented by thorough analysis, reveal that:

- LLMs, when subjected to fine-tuning, are potent MT models. Through fine-tuning, they consistently outperform their zero-shot prompting counterparts, achieving an average improvement of 8 BLEU for sentence-level translation and 16.33 BLEU for document-level translation. Notably, the model *opt-13b* even sees a remarkable boost of 28.93 BLEU (from 4.56 to 33.49).

- There is a large variation in the performance across different LLMs. LLaMA 2 consistently outperforms others for both prompting and fine-tuning. BLOOMZ, initially lagging behind in prompting, ascends to top-tier models after fine-tuning. However, some models, despite benefiting from fine-tuning, either match or fall short of the performance of models trained from scratch. It is also noteworthy that larger models don't invariably outshine their smaller counterparts.

- When prompted, LLMs demonstrate superior performance in sentence-level translation. However, the application of fine-tuning yields more substantial enhancements in document translation, as reflected by both the BLEU and COMET scores. Notably, LLaMA 2 surpasses its performance in sentence-level translation when trained on documents.

- QLoRA accelerates the fine-tuning process without compromising model performance. To attain an equivalent BLEU score, it necessitates 21 times less training time and reduces the trainable parameters by 1370-fold compared to conventional fine-tuning.

2 Related Work

2.1 LLM Applications

Leveraging LLMs across a spectrum of downstream natural language processing (NLP) tasks is now a prevailing approach. However, the optimal strategies for utilizing these models both effectively and efficiently remain an open question. Broadly speaking, there are three primary methods to build applications based on LLMs:

- **Zero-shot prompting.**[1] This involves querying LLMs with a prompt that hasn't been seen in the training data of the model. Such prompts typically provide specific task instructions along with the main query. Given the sensitivity of LLMs to the structure and content of prompts, careful prompt engineering is crucial to achieve optimal performance.

- **Few-shot learning.** Often referred to as in-context learning, few-shot learning is a technique where LLMs are provided with a handful of examples to guide their responses. Zero-shot prompting can be considered a subset of this, where no examples are given. In few-shot learning, these examples are integrated into the prompt template, serving as context to instruct the model on how to respond.

- **Fine-tuning.** The two methods above allow for task adaptation without the need for further training on the LLMs. In contrast, fine-tuning involves extending the training of the LLMs using additional, task-specific data. This is particularly beneficial when such tailored datasets are available.

Yang et al. (2023) survey the 'use cases' and 'no use cases' of LLMs for specific downstream tasks, considering the three aforementioned methods, and conclude that LLMs excel in most NLP tasks.

2.2 LLMs for MT

Recent literature has begun to explore the application of LLMs for MT, an area that remained relatively under-explored until now. Both Hendy et al. (2023) and Moslem et al. (2023) underscore the superiority of GPT3 (Brown et al., 2020), GPT3.5 and ChatGPT (Bawden and Yvon, 2023) in MT

[1]Throughout this paper, we refer to 'zero-shot prompting' simply as 'prompting'.

using prompting. However, the former also indicates that these models may not consistently outperform SOTA MT systems and commercial translators. In a comparative study, Zhu et al. (2023) experiment with various LLMs, including GLM-7.5B (Lin et al., 2021), OPT-175B (Zhang et al., 2022), BLOOMZ-7.1B (Muennighoff et al., 2022), and ChatGPT. Their findings suggest that while these decoder-only LLMs are competitive, they still lag behind when compared to the encoder-decoder-based multilingual language model NLLB (Costa-jussà et al., 2022). Briakou et al. (2023) studied the impact of LLM data on MT.

Prompting strategies for MT are studied by Vilar et al. (2023) for PaLM (Chowdhery et al., 2022) and Zhang et al. (2023) for GLM-130B (Zeng et al., 2022). They reveal several challenges associated with MT prompting, such as issues with copying, mistranslation of entities, and hallucination. These challenges are echoed by Bawden and Yvon (2023), which identify similar constraints with prompting on BLOOM (Scao et al., 2022). However, they show these limitations can be mitigated in a few-shot learning setting. Sia and Duh (2022) investigated a light-weight tuning method akin to prefix tuning (Li and Liang, 2021). Sia and Duh (2023) and Wang et al. (2023) expand the evaluation to document-level translation.

While prior studies have highlighted the potential of LLMs in MT, their focus has been primarily on in-context learning. A significant gap remains in the exploration of fine-tuning LLMs specifically for MT tasks. Additionally, there is an evident absence of research that provides a comprehensive comparison among prompting, few-shot learning, and fine-tuning methodologies. Recognizing this oversight, the primary objective of this paper is to address and bridge this research gap.

3 QLoRA

QLoRA (Dettmers et al., 2023) is an efficient fine-tuning approach that reduces the memory usage of training without compromising the 16-bit task performance. The approach involves quantizing a pre-trained model to 4-bit precision. Subsequently, a compact set of learnable Low-rank Adapter (LoRA, Hu et al. (2021)) weights are added, which can be tuned through backpropagation.

LoRA Motivated by the empirical findings of Li et al. (2018) and Aghajanyan et al. (2020), which suggest that LLMs possess a notably low intrinsic dimension for their parameters, LoRA hypothesizes a similar low intrinsic rank for weights during model adaptation. Thus, LoRA introduces a reparameterization aimed at reducing dimensions. Specifically, it employs a low-rank decomposition to represent the pretrained weights, resulting in newly-added adapter weight matrices, with the rank r anticipated to be considerably smaller than the original weight matrices' dimension. During fine-tuning, the pretrained weights are frozen, with only the newly incorporated adapter updated via back-propagation. A key observation is that as the rank r is reduced, there is a corresponding decrease in the number of adaptable parameters.

4 Experimental Setup

4.1 Datasets

In this study, we focus on the translation direction from French to English due to its significant demand for high-quality translation and the availability of substantial parallel data. Our fine-tuning set includes the commonly used Europarl (Koehn, 2005) and News Commentary dataset from WMT14[2]. The dev and test sets are the newstest2013 and newstest2014 datasets, respectively, from WMT14. These datasets are constructed from documents, thus enabling a natural evaluation of document-level translation. Table 1 summarizes the statistics of the datasets.

	#sents	#docs	avg.sents/doc
train	2,366,117	21,430	144
dev	3000	126	24
test	3003	169	18

Table 1: Dataset statistics.

4.2 Baseline

We compare the performance of systems built upon LLMs against an NMT model trained from scratch using the Amazon Sockeye framework (Hieber et al., 2022). The model architecture is a 12-layer transformer with a model size of 1024, 16 attention heads, and 4096 hidden units in the feed-forward layers. We employ byte pair encoding (BPE, Sennrich et al. (2016)) separately for each language, setting the number of BPE symbols to 30k for both languages. The model is trained with a batch size of 4096, an initial learning rate of 0.0002, and a

[2]https://www.statmt.org/wmt14/translation-task.html

Model	Release Time	Data	Size (B)
GPT-Neo (Black et al., 2021)	Mar, 2021	English-centric	1.3; 2.7
OPT (Zhang et al., 2022)	June, 2022	English-centric	1.3; 2.7; 6.7
LLaMA2 (Touvron et al., 2023)	July, 2023	English-centric	7; 13
XGLM (Lin et al., 2021)	Nov, 2022	Multilingual	1.7; 2.9; 4.5; 7.5
BLOOMZ (Muennighoff et al., 2022)	Nov, 2022	Multilingual	1.7; 3; 7.1

Table 2: Overview of evaluated LLMs.

plateau-reduce learning rate scheduler. Additionally, we apply a dropout and label smoothing of 0.1, use the Adam optimizer with a warm-up of 10k steps, and set the checkpoint interval to 4000. Training is halted if there is no improvement in performance on the dev set for 32 consecutive checkpoints. The model has 4 billion parameters and is trained on a single NVIDIA V100 with 32G GPU memory.

This is a relatively standard NMT model, devoid of advanced techniques such as back translation, knowledge distillation, or ensembling, which could potentially elevate the model to state-of-the-art performance (Kocmi et al., 2022). However, the primary objective of this study is to compare the efficacy of using an off-the-shelf machine translation toolkit, which is widely accessible and requires minimal effort for machine translation practitioners, against building MT systems using LLMs. Importantly, both methods demand similar levels of effort in development, making this a fair comparison to ascertain the most efficient approach for practitioners and researchers alike.

4.3 Pretrained LLMs

We investigate a varied collection of pretrained LLMs accessible on HuggingFace (Wolf et al., 2020), all based on the transformer architecture. This collection comprises five distinct LLMs, each trained on either English-centric or multilingual data and available in multiple versions with varying parameter sizes. This results in a comprehensive assortment of 15 models, with parameter sizes ranging from 1.3 billion to 13 billion. Table 2 summarizes the models included in our study.

- **GPT-Neo** - a GPT-2 (Radford et al.) like causal language model trained on the Pile dataset (Gao et al., 2020), an 825 GiB English corpus.
- **OPT** - a suite of causal language models, where the largest one, OPT-175B, exhibits performance comparable to GPT-3 (Brown et al., 2020).
- **LLAMA 2** - pretrained on 2 trillion tokens of English-centric data. We used a fine-tuned version of the model, referred to as *LLAMA 2-CHAT*. This fine-tuned version demonstrates superior performance compared to open-source chat models across a wide range of benchmarks.
- **XGLM** - a multilingual language model trained on a balanced corpus covering 30 diverse languages with 500B tokens. The XGLM 7.5B outperforms GPT-3 on the FLORES-101 (Goyal et al., 2022) machine translation benchmark in few-shot learning scenarios.
- **BLOOMZ** - a multilingual BLOOM model (Scao et al., 2022) fine-tuned with the xP3 dataset (Muennighoff et al., 2022), which consists of multilingual datasets with English prompts, totaling 95 GiB of text.

The selection of these models enables us to assess the impact of various factors on translation performance, including the type of model (English-centric vs. multilingual) and model size. Additionally, the chosen sizes reflect the computational resources typically available to research institutes with limited GPU resources, such as university labs. This consideration ensures that our findings are applicable and accessible to a broad range of machine translation researchers and practitioners.

4.4 Prompted Tuning

We fine-tune LLMs using examples that include specifically formatted prompts (French: [fr sent] English:) and their corresponding responses ([en sent]). The dev set is also formatted in the same way. This approach customizes the model for the French-English machine translation task.

Sentence-level Prompts The inputs at the sentence level are formatted as follows:

French: [fr sent] English: [en sent] <eos>

We append the special token <eos> at the end of each sample to regulate the length of the text generated by the model. Without this, LLMs tend

to generate text continuously until they reach a predetermined length limit.

Document-level Prompts We use the given document boundaries to concatenate parallel sentences into document-level sequences. These parallel documents comprise an equal number of sentences in both languages. Our goal is to ensure that the models generate the same number of output sentences per document as the number of input sentences provided, facilitating sentence-level evaluation. We adopt the document mark-up used in Junczys-Dowmunt (2019), incorporating symbols for document start (<BEG>) and end (<END>), as well as sentence separators (<SEP>). In instances where documents exceed our sentence limit of 10, we substitute the <END> symbol with a break symbol (<BRK>) and commence the subsequent sequence with a continuation symbol (<CNT>) instead of <BEG> . Below is an example of a document input:

French: <BEG> [fr sent1] <SEP> [fr sent2] <SEP><END> English: <BEG> [en sent1] <SEP> [en sent2] <SEP><END>

4.5 Fine-tuning Setup

We configure the learning rate to 2e-4 and employ the Adam optimizer for the training process. A batch size of 32 is used, and the evaluation is performed every 1000 steps. The fine-tuning process is halted if there is no improvement in the model's performance over 16 consecutive checkpoints. For the LoRA configurations, the rank for the low-rank approximation is set to 64, and the scaling factor for the low-rank adaptation is set to 32. The trainable parameters are limited to the self-attention layers of the model. Additionally, a dropout rate of 0.05 is applied in the LoRA layer. The model weights are quantized to 4-bit precision to reduce memory requirements, and mixed-precision training is enabled, using a combination of float16 and float32 data types to accelerate the training process. Models with less than 3 billion parameters are trained on a single NVIDIA RTX GPU with 24GB of memory, while models with more than 3 billion but less than 7 billion parameters are trained on a single NVIDIA V100 GPU with 32GB of memory. For models with an even larger number of parameters, we employ multiple V100 GPUs and enable model parallelism by setting device_map="auto" . This is facilitated by the Accelerate library from Hugging Face, which automatically distributes the model across the available GPUs.

4.6 Evaluation Metrics

We use **BLEU** and **COMET** (Rei et al., 2020) as evaluation metrics to assess the performance of our models. For BLEU we use the SacreBLEU (Post, 2018) implementation, which standardizes tokenization and facilitates reproducibility.

On the other hand, unlike BLEU, which depends on the n-gram overlap between the machine-generated translation and the reference translation, COMET models are trained on a comprehensive dataset comprising human translations and human quality assessments. This dataset is used to predict translation quality while also taking the source side into account. This approach enables COMET to provide a more holistic evaluation that includes fluency, adequacy, and preservation of meaning. We employ the latest model, *Unbabel/wmt22-comet-da*, for our evaluation. This model scales the scores between 0 and 1, where a score approaching 1 indicates a high-quality translation.

By employing both BLEU and COMET, we can ensure that our evaluation is robust and comprehensive, accounting for not only the lexical similarity between the translations and the references but also the overall quality and preservation of meaning in the translations. Moreover, COMET may serve as a superior evaluation metric when assessing the zero-shot performance of LLMs compared to BLEU. As we demonstrated in Section 7, the outputs from LLMs often excel in preserving meanings but might receive a low score if evaluated solely based on n-gram matching.

5 Sentence-level Translation

In this section, we assess the sentence-level translation performance of pretrained LLMs using prompting versus fine-tuned LLMs (Section 5.1). We investigate the effects of incorporating or not incorporating QLoRA during the fine-tuning process (Section 5.2). Additionally, we analyze the impact of varying QLoRA hyperparameters (Section 5.3), including the rank of the low-rank approximation (Section 5.3.1), and the trainable parameters (Section 5.3.2). We also conduct experiments with different sizes of fine-tuning data and compare the results of fine-tuned LLMs with the baseline NMT model (Section 5.4). Lastly, we explore few-shot learning with varying numbers of shots and diverse prompts (Section 5.5).

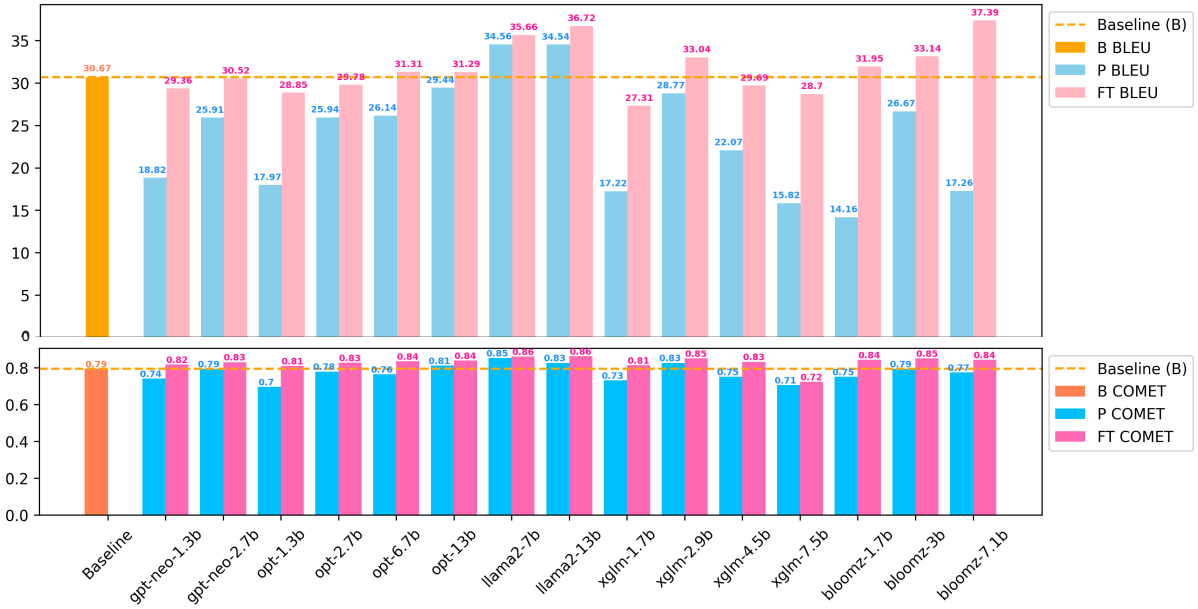


Figure 1: Prompting (*P*) vs. QLoRA fine-tuning (*FT*) on sentence-level translation using various pretrained LLMs. *Baseline* is the NMT system described in Section 4.2. Rank r for QLoRA is set to 64.[3]

5.1 Main Results

We present the results of prompting and QLoRA fine-tuning in Figure 1. Key observations are:

- While there is a significant disparity in BLEU scores, the same is not observed in COMET. All models exhibit comparable COMET scores. The top-performing fine-tuned model, *llama2-13b*, outperforms the *baseline* from 0.837 to 0.862. This indicates that while all models produce semantically coherent translations, their lexical choices, which affect BLEU scores, might differ.
- In terms of BLEU, the *baseline* model surpasses most prompted LLMs, with the exception of *LLAMA 2*. Specifically, *llama2-7b* achieves the highest performance at 34.56 BLEU, marking a 3.89 BLEU improvement over the *baseline*.
- 8 out of the 15 fine-tuned LLMs exceed the *baseline*. This includes both English-centric and multilingual models. The standout model is *bloomz-7.1b* achieving a BLEU score of 37.39, a 6.72 BLEU enhancement compared to the *baseline*.
- Fine-tuning invariably boosts LLM performance on average by 8 BLEU points, with *bloomz-7.1b* witnessing the most substantial leap of 20.13 BLEU.
- No clear advantage is discerned when contrasting prompted multilingual models with English-centric ones. For instance, the multilingual *bloomz-1.7b* scores the lowest at 14.16 BLEU. Yet, when evaluating the fine-tuning gains over prompting, multilingual models average an 11.32 BLEU improvement, surpassing the 5.02 BLEU of their counterparts.
- Bigger models do not consistently outshine their smaller counterparts. For instance, after fine-tuning, *bloomz-1.7b* trumps the larger *opt-13b* (31.95 vs. 31.29 BLEU). Within the same architecture, models with more parameters typically fare better, but there are exceptions, like with *XGLM*, where the *4.5b* and *7.5b* versions lag behind the *2.9b* variant.

In conclusion, while directly prompted LLMs do not universally outperform train-from-scratch MT models, certain LLMs, such as LLAMA 2, defy this trend. Moreover, fine-tuning consistently proves beneficial, with the potential to elevate even underperforming LLMs, like *bloomz-7.1b*, to top-tier performance.

	params(%)	**#GPUs**	**time(hrs)**
No QLoRA	27.40	4	52
QLoRA	0.02	1	10

Table 3: Fine-tuning *xglm-2.9b* with and without QLoRA to achieve the BLEU score of 30.05.[4]Only the self-attention layers are tuned. The rank r for QLoRA approximation is set to 2.

[3]We also report TER in Appendix A.

[4]We train the model without QLoRA for 96 hours in total, and 30.05 is the BLEU score obtained at the best checkpoint.

r	2	4	8	16	32	64	128	256	512
train params(%)	0.02	0.05	0.09	0.19	0.39	0.77	1.53	3.01	5.85
BLEU	31.69	31.72	32.28	32.52	32.80	**33.04**	30.60	30.09	30.31
COMET	0.845	0.846	0.847	0.848	0.849	**0.850**	0.837	0.835	0.836

Table 4: QLoRA fine-tuning results on XGLM 2.9B with various rank r choices. All the weights except for self-attentions are frozen.

5.2 QLoRA vs. No QLoRA

To assess QLoRA's efficacy, we contrast it with the original approach, a more resource-intensive choice: fine-tuning without QLoRA, which excludes both quantization and low-rank adaptation. We train the *xglm-2.9b* model using its native 32-bit precision, necessitating the use of 4 NVIDIA v100s. This is compared against a model fine-tuned with QLoRA set at $r = 2$. For consistency, only the self-attention layers are unfrozen in both models. The comparative results are presented in Table 3.

Achieving a BLEU score of 30.05, the model fine-tuned without QLoRA requires 52 hours across 4 GPUs, totaling 208 GPU hours. In contrast, the QLoRA-enhanced model completes in just 10 hours, marking a 21-fold acceleration and utilizing 1370 times fewer trainable parameters (0.02% compared to 27.4%).

5.3 QLoRA Hyperparameters

We investigate the impact of selecting different ranks for LoRA and the unfrozen parameters for fine-tuning. We present the results for XGLM 2.9B.

5.3.1 Rank r

The rank r of the decomposition matrices influences the number of trainable parameters, with a larger r resulting in more trainable parameters. We assess the performance associated with different choices of r, ranging from 2 to 512, in Table 4, while only unfreezing the self-attention layers.

With $r = 64$, the model attains its optimal performance. However, either reducing or increasing the number of trainable parameters adversely affects the model's performance. Interestingly, when $r = 512$, the performance deteriorates even more than when $r = 2$, despite the fact that the latter converges more quickly due to a smaller number of trainable parameters.

5.3.2 Trainable Parameters

Next, we aim to determine which part of the model should be fine-tuned. To do this, we unfreeze the parameters in different layers of the XGLM 2.9B model. As illustrated in Table 5, we experiment with unfreezing parameters from various layers, including the self-attention layers, embedding layers, fully-connected feed-forward layers, and the LM head layers. The results indicate that fine-tuning only the self-attention layer is sufficient to yield the best performance.

Params	a	a+e	a+e+f	a+e+f+l
BLEU	**31.69**	30.09	30.30	28.39
COMET	**0.845**	0.837	0.834	0.826

Table 5: QLoRA fine-tuning results on XGLM 2.9B with different trainable parameters. *a*: self-attentions; *e*: embeddings; *f*: fully-connected feed-forward layers; *l*: lm head. Rank r is set to 2.

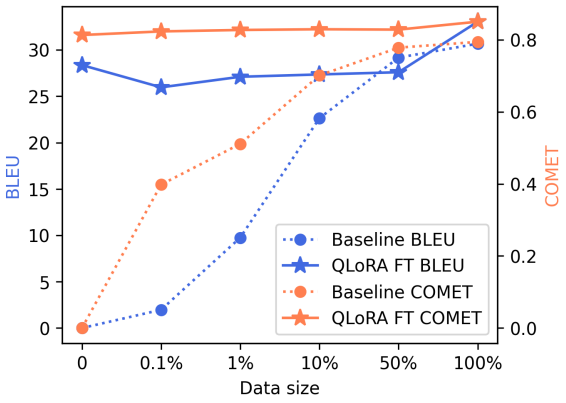


Figure 2: The performance of the *baseline* system and fine-tuned XGLM 2.9B trained with different amounts of data.

5.4 Data Curves

The performance of a traditional MT model is closely tied to the volume of its training data, as highlighted by (Koehn and Knowles, 2017). However, for LLMs, which have already benefited from vast training datasets, does this correlation still hold? To investigate, we compare the responses of both MT model types to varying training data sizes. We incrementally adjust the dataset size from 0.1% (2,366 examples) to its entirety and then train the

Prompt 1	{ **French:** [fr sent] **English:** [en sent] } x **K** **French:** [fr sent] **English:**
Prompt 2	{ **Translate French to English: French:** [fr sent] **English:** [en sent] } x **K** **Translate French to English: French:** [fr sent] **English:**
Prompt 3	**Translate French to English:** { **French:** [fr sent] **English:** [en sent] } x **K** **Translate French to English: French:** [fr sent] **English:**
Prompt 4	**Translate French to English:** **French:** { [fr sent] } x **K** **English:** { [en sent] } x **K** **Translate French to English: French:** [fr sent] **English:**
Prompt 5	{ **French:** [fr sent] **Translate to English:** [en sent] } x **K** **French:** [fr sent] **Translate to English:**

Table 6: Prompts used in **K**-shot learning. The substrings within {} are repeated **K** times.

	BLEU				**COMET**			
	0-shot	**1-shot**	**5-shot**	**10-shot**	**0-shot**	**1-shot**	**5-shot**	**10-shot**
Prompt 1	27.08	29.15	29.72	29.62	0.814	0.828	0.833	0.834
Prompt 2	28.36	29.46	29.86	29.95	0.813	0.830	0.836	0.835
Prompt 3	28.36	29.33	29.86	29.74	0.813	0.831	0.835	0.834
Prompt 4	28.36	29.46	28.66	27.83	0.813	0.830	0.829	0.825
Prompt 5	11.82	28.76	29.80	29.70	0.631	0.827	0.834	0.834

Table 7: Few-shot learning results on XGLM 2.9B.

baseline model and fine-tune the LLMs. The outcomes of this experiment are depicted in Figure 2.

The *baseline* curve validates the assumption that performance improves with increased data availability. In contrast, LLMs make a robust debut; even without additional training data, they achieve a BLEU score comparable to the *baseline* trained on half the dataset. Yet, their performance does not consistently improve with more data. In fact, fine-tuning with less than 50% (1.2 million examples) of the data seems counterproductive, diminishing performance until the full dataset comes into play.

5.5 Few-shot Learning

In this section, we evaluate the few-shot learning performance of LLMs. Few-shot learning is also denoted as **K**-shot, with **K** representing the number of examples provided before the query, where in our case, examples are randomly sampled from the training set. We also compare the impact of 5 slightly varied prompts, detailed in Table 6. The results of the experiments are presented in Table 7.

When **K** $>=$ 1, the model consistently outperforms the 0-shot scenario. For *prompt 5*, 1-shot dramatically enhances the model's capability, elevating the BLEU score from 11.82[5] to 28.76. However, the performance does not exhibit a linear growth with increasing **K**; it plateaus. In the case of *prompt 4*, augmenting **K** even diminishes the performance.

In our experiments, the choice of prompt is particularly impactful for 0-shot performance, especially when comparing *prompt 5* to the others. However, this impact seems to lessen when examples are presented before the query.

6 Document-level Translation

In this section, we delve into the proficiency of LLMs in document-level translation. Our primary observations, contrasting the prompted and fine-tuned LLMs, are detailed in Section 6.1. Additionally, we explore the influence of document length, measured by the number of sentences per document, in Section 6.2.

6.1 Main Results

Figure 3 presents the results for document-level translations. Key takeaways include:

[5]We observed many empty generations when prompting with **Prompt 5**. One hypothesis is that the prompt is ambiguous and the model is confused about what to translate.

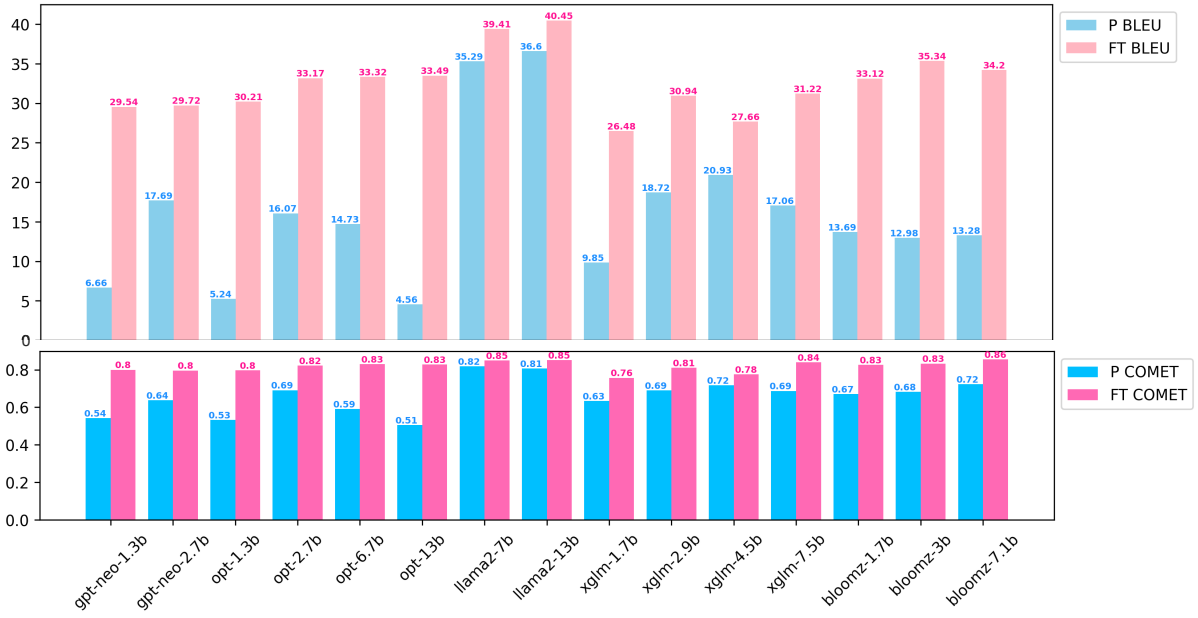


Figure 3: Prompting (*P*) vs. QLoRA fine-tuning (*FT*) on document-level translation using various pretrained LLMs. Rank r for QLoRA is set to 64.

- In contrast to sentence-level translation, prompted LLMs face challenges with document-level translation. 4 out of the 15 LLMs register BLEU scores below 10. However, consistent with sentence-level findings, *LLAMA 2* continues to stand out in zero-shot performance, with the *7b* and *13b* versions achieving impressive BLEU scores of 35.29 and 36.6, respectively.
- Fine-tuning demonstrates significant promise for document-level translations, enhancing the BLEU scores of their prompted counterparts by an average of 16.33. The most notable improvement is seen in *opt-13b*, which witnesses a BLEU increment of 28.93 (from 4.56 to 33.49).
- Unlike sentence-level translation, where COMET scores remain consistent across all models, document-level translation displays a more pronounced variance. This variability is particularly evident in prompted models but diminishes in fine-tuned ones.
- Trends observed in sentence-level translation (Section 5.1) persist in the document-level context: (1) Both English-centric and multilingual models deliver comparable performance. (2) Larger models do not consistently surpass their smaller counterparts. (3) Top-performing models in sentence-level translation, such as *LLAMA 2* and *BLOOMZ*, maintain their dominance in document-level translation.

6.2 Input Lengths

We construct the examples for fine-tuning by merging sentences from an original document, as described in Section 4.4. By varying the number of sentences per document – specifically, 5, 10, and 15 – we present our findings in Table 8. Notably, fine-tuning with sets of 10 consecutive sentences yields the optimal performance, registering a 30.94 BLEU and a 0.811 COMET score.

#sents	5	10	15
BLEU	29.07	**30.94**	28.32
COMET	0.781	**0.811**	0.742

Table 8: QLoRA fine-tuning on documents composed of different numbers of sentences[6]. XGLM 2.9B systems are fine-tuned with $r = 64$, and self-attentions and embeddings as trainable parameters.

7 Qualitative Analysis

Figure 4 shows translations of two French sentences using various LLMs. In the first example, when prompted, *bloomz-7.1b* replicates the source sentence verbatim, neglecting to translate. While this does not occur for every test set sample, as shown in the second example, similar behavior is noted in other prompted LLMs. On the other hand,

[6]We match the length of text documents to that of train documents.

French	L'ONU donne un bilan même plus élevé avec 979 morts et 1 902 blessés.
English	The UN has reported even higher numbers with 979 dead and 1,902 injured.
bloomz-7.1b P	L'ONU donne un bilan même plus élevé avec 979 morts et 1 902 blessés.
bloomz-7.1b FT	The UN gives a higher figure with 979 dead and 1 902 wounded.<eos>.<eos>.<eos>.
llama2-13b P	979 deaths and 1,902 injuries, according to the UN's latest tally.
llama2-13b FT	The UN gives an even higher death toll of 979 and 1 902 injured.<eos>The UN gives an even higher death toll of 979 and 1 902 injured.<eos>The UN gives an even higher death toll of 979 and 1 902 injured.<eos>The
French	L'affaire NSA souligne l'absence totale de débat sur le renseignement
English	NSA Affair Emphasizes Complete Lack of Debate on Intelligence
bloomz-7.1b P	French: The NSA case highlights the complete absence of debate on intelligence.
bloomz-7.1b FT	The NSA Case Sheds Light on the Absence of Debate on Intelligence<eos>.<eos>.<eos>.
llama2-13b P	The NSA case highlights the complete lack of debate on intelligence gathering.
llama2-13b FT	The NSA Scandal Highlights the Lack of Intelligence Debate<eos>eos>eos>

Figure 4: Translations from prompted (**P**) and fine-tuned (**FT**) LLMs.

the translation using *llama2-13b P*, though not mirroring the reference verbatim, retains the original sentence's meaning. Both fine-tuned LLMs produce proper translations with the initial segment of the generated sequences. *Bloomz-7.1b* appends a <eos> token post-translation, while *llama2b-13b* reiterates its translation multiple times. Both outputs necessitate post-processing, specifically truncating the output at the first occurrence of the <eos> token.

In the second example, the LLM-generated translations retain the meaning of the reference translation, showcasing LLMs' potential in the translation tasks.

8 Conclusions

In this study, we investigate the capabilities of LLMs in performing machine translation tasks. Through comprehensive experiments, we assess the effectiveness of prompting, few-shot learning, and fine-tuning using QLoRA for French-English translation. Our key findings are:

1. The proficiency of LLMs in machine translation varies. While **LLAMA 2** consistently outperforms its counterparts, other models, when relying solely on few-shot learning, often lag behind models trained from scratch.
2. Fine-tuning invariably enhances performance, particularly for models that struggle with few-shot learning and for translating documents. It can transform a seemingly inadequate model into a top-tier translation model, as seen with *bloomz-7.1b*.
3. QLoRA, due to its efficiency, can be a superior alternative to original fine-tuning methods.
4. Fine-tuning LLMs with QLoRA can be a promising and new paradigm for machine translation practice.

In the future, we are interested in exploring two primary avenues. (1) While our current study demonstrates the promise of LLMs trained on English-centric data for French-to-English translations, it raises intriguing questions: Would these results hold true for other language pairs, especially for low-resource languages? And would there be a noticeable difference in performance between English-centric and multilingual LLMs in such scenarios? (2) Our experiments are confined to decoder-based LLMs. Moving forward, we are also interested in comparing these models against their encoder-decoder counterparts, such as mT5(Xue et al., 2021), mBART (Liu et al., 2020), NLLB (Costa-jussà et al., 2022).

Limitations

Single dataset and language pair Our experiments are confined to a single dataset and the French-English language pair. It remains unclear if our findings are generalizable to other datasets and language pairs.

Medium-sized LLMs We have only experimented with medium-sized LLMs due to computational resource constraints. The necessity of fine-tuning for significantly larger LLMs remains an open question.

Acknowledgements

This work is supported in part by an Amazon Initiative for Artificial Intelligence (AI2AI) Faculty Research Award.

References

Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. 2020. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of bloom. *arXiv preprint arXiv:2303.01911*.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for needles in a haystack: On the role of incidental bilingualism in PaLM's translation capability. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9432–9452, Toronto, Canada. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Felix Hieber, Michael Denkowski, Tobias Domhan, Barbara Darques Barros, Celina Dong Ye, Xing Niu, Cuong Hoang, Ke Tran, Benjamin Hsu, Maria Nadejde, et al. 2022. Sockeye 3: Fast neural machine translation with pytorch. *arXiv preprint arXiv:2207.05851*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at wmt 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 225–233.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. 2018. Measuring the intrinsic dimension of objective landscapes. *arXiv preprint arXiv:1804.08838*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yasmin Moslem, Rejwanul Haque, and Andy Way. 2023. Adaptive machine translation with large language models. *arXiv preprint arXiv:2301.13294*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *ICLR 2022-Tenth International Conference on Learning Representations*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725. Association for Computational Linguistics (ACL).

Suzanna Sia and Kevin Duh. 2022. Prefix embeddings for in-context machine translation. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 45–57, Orlando, USA. Association for Machine Translation in the Americas.

Suzanna Sia and Kevin Duh. 2023. In-context learning as maintaining coherency: A study of on-the-fly machine translation using large language models. In *Proceedings of Machine Translation Summit XIV (Volume 1: Research Track)*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Xin Tan, Longyin Zhang, Deyi Xiong, and Guodong Zhou. 2019. Hierarchical modeling of global context for document-level neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1576–1585.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2018. Learning to remember translation history with a continuous cache. *Transactions of the Association for Computational Linguistics*, 6:407–420.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. *arXiv preprint arXiv:2304.02210*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Hongfei Xu, Deyi Xiong, Josef Van Genabith, and Qiuhui Liu. 2021. Efficient context-aware neural machine translation with layer-wise weighting and input-aware gating. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3933–3940.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. 2023. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. *arXiv preprint arXiv:2301.07069*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

A TER on Sentence-level Translations

The Translation Edit Rate (TER) is a metric introduced by Snover et al. (2006) to quantify the amount of human editing required to align a system's output with a reference translation. Specifically, TER is calculated as the ratio of the total edits made to the length of the reference translation. Such edits encompass insertions, deletions, single-word substitutions, and shifts in word sequence. A lower TER indicates better alignment with the reference. As illustrated in Figure 5, when evaluated using TER, LLMs do not exhibit a noticeable improvement over the baseline model.

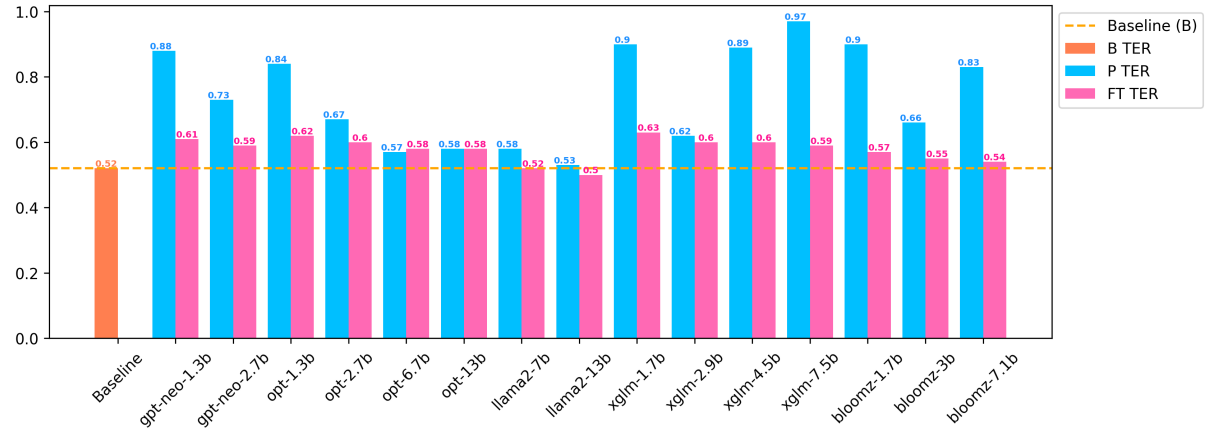


Figure 5: Prompting (*P*) vs. QLoRA fine-tuning (*FT*) on sentence-level translation using various pretrained LLMs. *Baseline* is the NMT system described in Section 4.2. Rank r for QLoRA is set to 64.

Towards Effective Disambiguation for Machine Translation with Large Language Models

Vivek Iyer **Pinzhen Chen** **Alexandra Birch**
School of Informatics, University of Edinburgh
{vivek.iyer, pinzhen.chen, a.birch}@ed.ac.uk

Abstract

Resolving semantic ambiguity has long been recognised as a central challenge in the field of Machine Translation. Recent work on benchmarking translation performance on ambiguous sentences has exposed the limitations of conventional Neural Machine Translation (NMT) systems, which fail to handle many such cases. Large language models (LLMs) have emerged as a promising alternative, demonstrating comparable performance to traditional NMT models while introducing new paradigms for controlling the target outputs. In this paper, we study the capabilities of LLMs to translate "ambiguous sentences" - i.e. those containing highly polysemous words and/or rare word senses. We also propose two ways to improve their disambiguation capabilities, through a) in-context learning and b) fine-tuning on carefully curated ambiguous datasets. Experiments show that our methods can match or outperform state-of-the-art systems such as DeepL and NLLB in four out of five language directions. Our research provides valuable insights into effectively adapting LLMs to become better disambiguators during Machine Translation. We release our curated disambiguation corpora and resources at https://data.statmt.org/ambiguous-europarl.

1 Introduction

While the field of NMT has advanced rapidly in recent times, the disambiguation and translation of ambiguous words still remain an open challenge. Notably, Campolungo et al. (2022) created a benchmark named DiBiMT to study the behaviour of state-of-the-art (SOTA) NMT systems when translating sentences with ambiguous words.[1] They reported that even the best-performing commercial NMT systems yielded accurate translations only 50-60% of the time,[2] while other open-source multilingual models like mBART50 (Tang et al., 2021) and M2M100 (Fan et al., 2021) performed much worse. This was found to be due to biases against rare and polysemous word senses inherited during pretraining. Table 1 shows an example from the DiBiMT benchmark where DeepL[3] mistranslates an ambiguous word while the LLM BLOOMZ resolves the word to its correct in-context meaning.

Source	The horse had a blaze between its eyes.
DeepL	那匹马的两眼之间有一团火焰。 (There is a flame between the horse's eyes.)
BLOOMZ (176B)	这匹马的眼睛之间有一道白线。 (There is a white line between the horse's eyes.)

Table 1: An example of English-to-Chinese translation involving an ambiguous term "blaze". For BLOOMZ, we use 1-shot prompting to obtain the translation.

In this paper, we explore whether LLMs can indeed perform better at translating "ambiguous sentences" – i.e. those containing highly polysemous and/or rare word senses. The motivation behind this is that while NMT models can potentially learn biases from noisy or narrow domain parallel data, hurting their ability to detect and translate rare word senses, LLMs can potentially be pretrained on a wider variety of monolingual text – though they might also prefer fluency over accuracy. Still, LLMs have shown many emergent abilities due to scale (Brown et al., 2020; Chowdhery et al., 2022; Wei et al., 2022a) and moreover, have demonstrated great potential for Machine Translation (MT) (Vilar et al., 2023; Zhang et al., 2023).

We comprehensively examine how these trends extend to the specific task of translating ambiguous sentences. We select a diverse set of foundational and instruction-tuned LLMs, of different

[1] https://nlp.uniroma1.it/dibimt/public/leaderboard

[2] Subsequent iterations of these commercial models have improved, but large margins still remain.

[3] https://deepl.com/en/translator

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 482–495
December 6–7, 2023.

sizes and with varying combinations of languages in the pre-training data. We then compare how these LLMs match up against several widely used NMT models on the DiBiMT test set, which covers translation from English to five languages: Spanish, Italian, German, Russian and Chinese. We find that, with only 1-shot in-context learning (Brown et al., 2020), LLMs – in particular, BLOOMZ 176B (Muennighoff et al., 2023) and LLaMA 65B (Touvron et al., 2023) – match or outperform top-performing open-source and commercial MT systems, and set a new SOTA in two of the five languages we tested. Furthermore, we propose two methods for adapting LLMs for ambiguous translation: 1) in-context learning with sentences having the same word sense, and 2) fine-tuning on curated ambiguous parallel corpora. We show that these methods are highly effective and can further improve performance by up to 15 points in DiBiMT accuracy in the best case.

Our work thus makes three key contributions:

1. We evaluate the performance of LLMs compared to top-performing NMT systems in the challenging task of translating ambiguous sentences. We report SOTA scores on 2 of the 5 languages tested, and comparable performance otherwise.

2. We also show that our suggested techniques of similar sentence in-context learning and targeted disambiguation fine-tuning significantly outperform naive few-shot prompting

3. We conclude our work by evaluating LLMs on the FLORES200 test sets, and confirm that improvements in disambiguation accuracy correlate strongly with those in overall MT quality.

2 Background

2.1 Ambiguity in machine translation

Resolving ambiguity in the source sentence was historically framed as one of the most fundamental challenges in MT (Weaver, 1952). In an effort to address this challenge, traditional works integrating Word Sense Disambiguation in Statistical Machine Translation (Carpuat and Wu, 2007; Chan et al., 2007) were followed by those integrating it in NMT architectures in various ad-hoc ways (Choi et al., 2017; Liu et al., 2018; Pu et al., 2018). Later, with the introduction of the Transformer (Vaswani et al., 2017), it was shown that higher layer encoder representations are robust enough to handle disambiguation (Tang et al., 2019) without any explicit handling of word senses.

However, more recent research creating challenging evaluation benchmarks has called the purported abilities of NMT systems into question once again. Following the proposal of the MuCoW benchmark for testing WMT19 (Raganato et al., 2019) and WMT20 (Scherrer et al., 2020) systems, Raganato et al. (2020a) showed how Transformer-based NMT models, in general, underperform when translating rare word senses. Campolungo et al. (2022), who experimented with SOTA commercial (Google Translate, DeepL) and open-source systems (mBART50, M2M100, OPUS-NMT (Tiedemann and Thottingal, 2020), etc.), arrived at the same conclusion when they proposed the DiBiMT benchmark for evaluating MT systems between English and 5 languages (Spanish, Italian, German, Russian, and Chinese). They found similar biases against low-frequency and highly polysemous word senses. They also noted the accuracies of these systems were much lower than the then SOTA WSD system, ESCHER (Barba et al., 2021) – indicating significant room for improvement. In this work, we explored whether foundational and instruction-tuned LLMs could bridge this gap with minimal supervision (i.e. few-shot prompting).

2.2 LLMs and machine translation

Previous research has found that LLMs can perform machine translation without being specifically fine-tuned (Radford et al., 2019). In order to elicit a translation, research in this direction follows the paradigm of LLM prompting:

1. Zero-shot prompting, where an LLM is directly asked to translate a source input into the target language (Radford et al., 2019).

2. Few-shot prompting, also called in-context learning, where an LLM is supplied with demonstrations of input and output pairs from the same task it is performing, before being queried an input (Brown et al., 2020).

3. Chain-of-thought (CoT), where an LLM is prompted to reason to gain relevant knowledge about the input before producing an output (Wei et al., 2022b; Kojima et al., 2022).

Besides training-free approaches, another route is instruction tuning, which optimizes an LLM on a

mixed range of downstream tasks and fine-tunes the model to understand and respond to user intention through natural language (Wei et al., 2021).

It was observed that LLMs might not surpass Transformer models solely trained to translate, especially for non-English and low-resource translation directions (Vilar et al., 2023; Hendy et al., 2023). Nevertheless, LLMs have been shown to achieve superiority in tasks requiring in-depth understanding and manipulation of text, primarily due to them being pretrained on very large corpora. For example, without fine-tuning, LLMs are good at adapting to word alignments (Moslem et al., 2023), translation evaluation (Kocmi and Federmann, 2023), idiom translation (Raunak et al., 2023), iterative refinement (Chen et al., 2023), and interactive translation via CoT (Pilault et al., 2023; He et al., 2023). Related to our work is Pilault et al. (2023)'s proposal of using interactive question answering as a CoT process for LLMs to disambiguate source words. As an alternative approach, we aim to generate translations in a single pass by leveraging SOTA WSD systems to provide contexts that guide LLMs to disambiguate better.

3 Methodology

3.1 Preliminaries

A word sense is a concept in a Knowledge Base (in this work, BabelNet by Navigli et al. (2021)) that denotes a distinct meaning of a word in the context of a sentence. The polysemy degree of an ambiguous word is defined as the total count of all possible senses that a particular word can have. The sense frequency is defined as the occurrence count of that particular sense in a disambiguated training corpus.

In this work, we define an ambiguous word as a polysemous term with multiple possible, and likely related, meanings – with the correct sense inferable only from the sentence-level context. We then refer to a sentence with an ambiguous word as an "ambiguous sentence" for brevity and ease of explanation. By definition, the DiBiMT test set (Campolungo et al., 2022) contains only one ambiguous word per sentence.

Word Sense Disambiguation (WSD) is the process of linking an ambiguous word in a sentence to its appropriate word sense in the Knowledge Base. We use ESCHER-WSD (Barba et al., 2021) in this work, a high-performing WSD system that had achieved the SOTA for English.

3.2 *K*-shot prompting

Given a test sentence X and a Large Language Model to prompt for translations, we construct a query with k demonstrations, i.e. parallel sentence pairs $\{(X_1, Y_1), (X_2, Y_2) \dots (X_k, Y_k)\}$ as examples, followed by the test sentence. As shown in Figure 1, for foundation LLMs, we frame the prompt as a text completion task, while for instruction-tuned LLMs (like BLOOMZ) we structure the last phrase as a question, in order to conform to the latter's question answering format. In the naive setting, we choose our demonstrations randomly from the development set.

3.3 In-context learning with similar ambiguous contexts

LLMs can effectively gain knowledge relevant to the test domain through prompting, and this process is named in-context learning (ICL). We leverage ICL to help LLMs ingest information on translation of ambiguous sentences, by providing related sense translations as examples in the prompt. To achieve this, we first identify the most polysemous word in the input sentence by disambiguating it with a WSD system, and then calculate the polysemy degree of all disambiguated senses with respect to a large development set. We choose the most polysemous word sense[4] and search for other occurrences of the same sense in the same development set. Finally, we randomly sample k source-target pairs including such a sense to use as demonstrations in k-shot prompting, instead of using random pairs. This technique seemed to return enough examples for our purposes in most cases – for 5-shot prompting, given a corpus of 1.8M sentences, we observed that we got all 5 matches 92.5% of the time.

3.4 Low-rank fine-tuning

Apart from providing relevant examples through prompting, another conventional approach is to optimize the model parameters in a domain adaptation fashion for disambiguation. Considering the computational cost, our work experiments with instruction fine-tuning via low-rank adaptation (LoRA). This technique appends trainable lower-rank decomposition matrices to giant matrices in an LLM

[4]Currently, we only explore the case of one ambiguous word per sentence, due to the nature of the benchmark. One could extend our approach to multiple ambiguous words by separately sampling examples for each polysemous word and conducting higher-shot prompting - but further research would be needed to find the optimal way to combine these examples.

```
Translate the following sentence from {src_lang} to {tgt_lang}: {src_demo_1}
The translation in {tgt_lang} is: {tgt_demo_1}
                              ⋮
                              k demonstrations
                              ⋮
Translate the following sentence from {src_lang} to {tgt_lang}: {src_demo_k}
The translation in {tgt_lang} is: {tgt_demo_k}
[FOUNDATION LLM]
Translate the following sentence from {src_lang} to {tgt_lang}: {src_test}
The translation in {tgt_lang} is:
                                    OR
[INSTRUCTION-TUNED LLM]
Translate the following sentence from {src_lang} to {tgt_lang}: {src_test}
Can you translate the input sentence to {tgt_lang}?
```

Figure 1: Templates used for k-shot LLM prompting, with $k >= 0$.

that can remain frozen during fine-tuning (Hu et al., 2021). By sacrificing a little performance, this fine-tuning method achieves great parameter efficiency. We aim to adjust LLMs to perform the translation task specifically. In order to maximise an LLM's capability to disambiguate when translating, we follow a careful data selection procedure to identify the most ambiguous sentences in our corpus.

Given the size of LLMs, it would be infeasible to fine-tune them on a large parallel corpus, so we opt to curate a smaller dataset that suits the ambiguous translation task. We would like a balanced mix of sentences with highly polysemous words as well as those with rare senses of a given word. This is to ensure fine-tuning reduces both polysemy degree-related and sense frequency-related biases, as discovered by Campolungo et al. (2022) and consequently, maximises disambiguation performance. We, thus, sort our corpora in two ways: one, by the maximum polysemy degree (greatest first) and two, by the minimum sense frequency (rarest first) of all word senses in a given sentence, disambiguated with ESCHER-WSD. We take the top $N/2$ sentences from each set and interleave them to create our final fine-tuning corpus of size N. We release our fine-tuning corpus, along with the ESCHER-WSD disambiguation outputs for public use.[5]

Once the data is chosen, we follow the fine-tuning paradigm of Alpaca (Taori et al., 2023): the model is prompted with an instruction specifying the source and target languages, as well as the test sentence as an input, and the model is expected to respond with the translation.[6]

[5] https://data.statmt.org/ambiguous-europarl

[6] https://github.com/tatsu-lab/stanford_alpaca

4 Experiments

In this section, we seek to answer the following research questions:

1. **RQ1:** How do LLMs perform at translation of ambiguous sentences compared to traditional high-performing NMT systems? (Section 4.3)
2. **RQ2:** What methods could one use to adapt LLMs for this task and improve performance over naive few-shot prompting? (Section 4.4)
3. **RQ3:** How do these disambiguation-adapted LLMs fare in terms of overall translation quality? (Section 4.5)

4.1 Models

To ensure reproducibility, we pick four well-known and high-performing open-source LLMs,[7] of which we sample seven versions for experimentation:

- BLOOM (Scao et al., 2022): A fully open-source, multilingual, foundation LLM that supports 46 languages. To establish the range of its capabilities, we explore both the smallest (7.1B) and the largest (176B) versions.
- BLOOMZ (Muennighoff et al., 2023): BLOOM instruction-tuned on a multilingual prompting set. Again, we choose the smallest (7.1B) and the largest (176B) versions.
- LLaMA (Touvron et al., 2023): The popular LLM trained by Meta AI, on gigantic datasets ranging up to 1.5T tokens. We evaluate the smallest (7B) and the largest (65B) versions.

[7] at the time of experiment formulation

- Alpaca (Taori et al., 2023): A LLaMA model instruction-tuned on a 52K dataset generated using Self-Instruct (Wang et al., 2023).

To effectively position these open-source LLMs against traditional NMT systems, we compare them against the best-performing and the most widely used commercial and open-source models:

1. DeepL Translator[8]: a SOTA commercial NMT system (accessed on 24th July 2023).
2. Google Translate[9]: Probably the most widely used commercial NMT system (accessed on 24th July 2023).
3. OPUS (Tiedemann and Thottingal, 2020): Small, bilingual, Transformer-based NMT models trained on the OPUS parallel corpora.
4. mBART50 (Tang et al., 2021): Multilingual NMT models pretrained on monolingual corpora from 50 languages, and fine-tuned on the translation task. We report performances of both the English-to-many and many-to-many fine-tuned models.
5. M2M100 (Fan et al., 2021): A massive multilingual NMT model that was trained on 2200 translation directions to support many-to-many translation among 100 languages in total. We compare both the base (418M) and the large (1.2B) versions.
6. NLLB-200 (NLLB Team et al., 2022): It is the current SOTA in many low-resource pairs, scaling to 200 languages. We experiment with all its variants, where the largest is a mixture-of-experts (MoE) model with 54B parameters. We also benchmark its smaller checkpoints at 1.3B and 3.3B, as well as distilled versions at 0.6B and 1.3B.

We take the results for mBART50, M2M100, and OPUS directly from the DiBiMT leaderboard.[10] We use Hugging Face[11] for accessing and inferencing all other models – except for Google Translate and DeepL, which are accessed using their respective APIs. Despite their presence on the leaderboard, we re-evaluate these systems since they are being constantly updated.

[8] https://www.deepl.com/en/translator
[9] https://translate.google.com/
[10] https://nlp.uniroma1.it/dibimt/public/leaderboard
[11] https://huggingface.co/

System	En-Es	En-It
Similar contexts dev set	1.81M	1.73M
Fine-tuning corpus	100K	100K

Table 2: Statistics of data used in our experiments, in terms of parallel sentence count.

4.2 Experimental setup

Datasets In this study, we use the DiBiMT test set for evaluation and measure accuracy across all five translation directions: English to Spanish, Italian, Chinese, Russian, and German, respectively. For validation, we use the development set from FLORES 200 (NLLB Team et al., 2022) in our base setting. To search for similar ambiguous contexts (Section 3.3), we require a larger development set to find relevant examples and also to accurately estimate polysemy degree. Hence, we use the Europarl corpus (Koehn, 2005), disambiguated with ESCHER-WSD. We also use the same disambiguated corpus for fine-tuning, however, we first follow the filtering procedure described in Section 3.4 to create a small corpus full of ambiguous sentences. Validation during fine-tuning is done using 500 randomly sampled sentences from this corpus and the rest is used for training. We detail the data statistics used for these experiments in Table 2.

LLM prompting setup Due to memory constraints, and to compare all models fairly, we load LLMs in 8-bit and use a batch size of 1. For generation, we set both beam size and temperature to 1. To prevent repetition in LLM output, we set `no_repeat_ngram_size` to 4. From the LLM's response, we filter out the sentence before the first newline character as the output translation.

LoRA fine-tuning We inject LoRA modules into all query, key, and value matrices. We set rank to 8, alpha to 8, and dropout to 0.05. For training, we set the effective batch size to 32, the learning rate to 3e-4, and the maximum length to 256. The total training budget is 5 epochs, and we pick the best model checkpoint based on cross-entropy loss on the validation set. The training data is shuffled after every epoch. Inference is done with a beam size of 3, and a maximum generation length of 150.

4.3 LLMs vs NMT systems on DiBiMT

We show our results in Table 3. For the subsequent discussion, we note that LLaMA was not intentionally trained on Chinese and is, thus, an 'unseen'

System	# Params	Variant	En-Es	En-It	En-Zh	En-Ru	En-De	Average
Commercial systems								
DeepL	Unknown	July 2023	63.91	**65.47**	58.42	**67.53**	**76.64**	**66.39**
Google Translate	Unknown	July 2023	54.73	53.59	52.09	62.03	***67.35***	57.96
Open-source NMT systems								
OPUS	74M	Bilingual En-X models	36.79	29.93	25.94	28.71	27.04	29.68
mBART50	611M	One-to-Many	31.31	26.62	26.63	30.93	26.43	28.38
	611M	Many-to-Many	29.98	25.89	28.12	27.54	24.25	27.16
M2M100	418M	Base	22.35	17.27	12.34	17.01	15.62	16.92
	1.2B	Large	28.81	23.16	17.30	27.03	22.87	23.83
NLLB-200	0.6B	Distilled version	40.93	36.38	28.64	47.13	33.41	37.30
	1.3B	Distilled version	50.40	53.65	41.15	54.52	52.81	50.51
	1.3B	Original checkpoint	48.81	48.43	37.31	54.36	48.93	47.57
	3.3B	Original checkpoint	53.23	57.23	39.95	57.44	56.24	52.82
	54B	Mixture of Experts	61.33	**67.19**	48.02	**67.88**	**67.97**	**62.48**
LLaMA family LLMs								
LLaMA	7B	1-shot prompting	53.64	48.84	30.61†	60.65	57.41	50.23
		3-shot prompting	55.53	50.53	30.52†	57.31	55.34	49.85
		5-shot prompting	56.33	48.66	27.92†	56.83	55.26	49.00
	65B	1-shot prompting	56.57	60.22	44.73†	65.71	62.05	57.86
		3-shot prompting	59.83	60.18	42.77†	***67.45***	63.41	58.73
		5-shot prompting	60.78	***63.47***	42.49†	66.31	62.98	***59.21***
Alpaca	7B	0-shot prompting	49.75	45.24	29.63†	55.23	51.52	46.27
BLOOM family LLMs								
BLOOM	7.1B	1-shot prompting	55.69	28.79†	51.08	40.00†	29.67†	41.05
	176B	1-shot prompting	63.66	42.02†	60.30	43.22†	37.04†	49.25
		3-shot prompting	64.52	46.33†	61.20	44.30†	36.69†	50.61
		5-shot prompting	65.53	45.99†	61.73	42.92†	38.06†	50.85
BLOOMZ	7.1B	0-shot prompting	56.89	33.91†	53.2	33.33†	21.67†	39.80
		1-shot prompting	60.87	40.68†	52.37	33.33†	30.65†	43.58
	176B	0-shot prompting	62.67	45.78†	61.87	47.98†	44.06†	52.47
		1-shot prompting	***64.35***	49.31†	**66.57**	51.88†	43.92†	55.21
		3-shot prompting	**67.31**	45.91†	**64.44**	53.42†	45.08†	55.23
		5-shot prompting	**68.55**	49.22†	***63.36***	52.60†	44.94†	55.73

Table 3: Accuracies on DiBiMT test for establish NMT systems and LLMs, using naive k-shot prompting. For Alpaca, we can only use 0-shot prompting due to its particular prompt template. We highlight the top three scores per language in bold, with the best underlined as well, the 2nd best as is, and the 3rd best italicized. We indicate scores for unseen languages (ie. not intentionally included in pretraining) with a †.

language. Similarly, for BLOOM, Chinese and Spanish are "seen" and the rest are "unseen". We share our key observations below:

1. **LLMs usually match or beat massive MT models on seen languages.** Except for the very rich-resourced En-De, where supervised MT systems appear to have an edge, LLaMA 65B mostly matches the SOTA NMT systems (namely DeepL and NLLB-200). Furthermore, BLOOMZ sets a new SOTA in its seen languages, Spanish and Chinese, and outperforms DeepL by margins of 7.3% and 12.2% respectively. These improvements against such strong, supervised massive NMT systems are particularly remarkable since our corresponding setup for inferencing the LLMs is quite cheap – as we noted previously, this is only naive few-shot prompting of an 8-bit quantized model, with a beam size of 1.

2. **LLMs perform relatively worse for unseen languages, but they can still be much better than some supervised MT models.** We note that relative to seen languages, LLaMA underperforms in translation to Chinese. Similarly,

BLOOM performs worse for its' unseen languages of German, Italian, and Russian. Still, LLMs yield reasonable performance here that is still much better than some supervised NMT systems. For example, BLOOMZ-7B achieves 40.68% accuracy in English-Italian, which is about 35.9% more than OPUS, 52.8% more than mBART50 and 75% more than M2M100-1.2B. While NLLB-200 does outperform BLOOMZ-7B, our results just highlight the power of pretraining at scale.

3. **Scale helps improve performance for ambiguity translation**. Continuing from the last point, similar to NMT models that improve with scale (e.g. NLLB-200), we observe that LLMs too perform consistently better at ambiguous translation on scaling up to their larger variants. This applies to the translation of both seen and unseen languages. That said, the lighter models, such as LLaMA 7B or BLOOM 7B, also perform quite well and in many cases, 1-shot prompting of these LLMs is almost as good as NLLB translations.

4. **LLM performance does improve on average with more demonstrations, but this is not uniform.** On average, we observe that 5-shot prompting works best, followed by 3-shot and then 1-shot, though some outliers exist for LLaMA 7B. Moreover, when looking at the performance of individual language pairs, we note that the improvement trend is not uniform, and it is possible a 3-shot translation outperforms a 5-shot one. This aligns with the finding of Zhang et al. (2023), who reach the same conclusion regarding overall MT quality. Nonetheless, as we show in Section 4.4.1, accuracy does significantly improve when we provide relevant and helpful examples – suggesting quality of demonstrations matters more than quantity.

5. **General-purpose instruction-tuned LLMs consistently outperform foundation LLMs.** Interestingly, we observe that 1-shot prompting of a general-purpose instruction-tuned LLM like BLOOMZ often significantly outperforms 5-shot prompting of BLOOM, even on the very specific task of ambiguity translation. In fact, even with 0-shot prompting, models like Alpaca 7B, BLOOMZ 7B and BLOOMZ 176B perform reasonably well, matching some supervised MT systems. We observed that this did not work for foundation LLMs like BLOOM 165B and LLaMA 7B, and 0-shot prompting of these models yielded hallucinations in many cases.

Lastly, we include a qualitative comparison of DeepL and BLOOMZ 176B translations for the En-Zh pair in the Appendix (see Table 8) – where we observe that BLOOMZ generates more contextual translations, relatively speaking, while its counterpart tends to translate literally in many cases.

4.4 Adapting LLMs for ambiguous MT

This section reports experiments with two proposed strategies to enable LLMs to disambiguate better and improve performance on the ambiguous translation task. While both methods are shown to significantly improve performance, we include a discussion of the relative tradeoffs between the techniques in Appendix A.2.

4.4.1 Improving In-Context Learning by leveraging similar ambiguous contexts

Rather than selecting our examples randomly as in our naive setting, we employ the data selection procedure described in Section 3.3 to discover other examples that contain the same word sense as the most polysemous sense in the input sentence. We report our scores in Table 4, and our findings below:

1. **Similar contexts yield more improvements as the example count increases** We observe that for 1-shot prompting, similar contexts perform comparably or slightly better than random examples. However, the gains increase substantially as we move towards 3-shot and 5-shot prompting. We can understand this from the intuition that 1-shot prompting likely just guides the LLM towards generating a reasonable translation, whereas with more relevant examples, it learns to disambiguate better and translate in context accordingly.

2. **Larger models observe greater and more consistent gains than smaller LLMs** Compared to LLaMA 7B, the other LLMs (LLaMA 65B, BLOOM 176B and BLOOMZ 176B) yield much larger accuracy improvements on a more uniform basis. This is probably because scaling up allows LLMs to model polysemous words better in their semantic space, facilitating effective in-context learning of disambiguation capabilities.

System	1-shot		3-shot		5-shot	
	Rand.	Sim.	Rand.	Sim.	Rand.	Sim.
DeepL	—63.91—					
NLLB-200 54B	—***61.33***—					
LLaMA 7B	53.64	**54.01**	**55.53**	52.52	**56.33**	54.45
LLaMA 65B	56.57	**59.38**	59.83	**62.44**	60.78	**63.74**
BLOOM 176B	**63.66**	62.44	64.52	**66.19**	65.53	**68.22**
BLOOMZ 176B	64.35	**69.57**	67.31	**71.15**	68.55	**71.33**

(a) English-Spanish

System	1-shot		3-shot		5-shot	
	Rand.	Sim.	Rand.	Sim.	Rand.	Sim.
DeepL	—65.47—					
NLLB-200 54B	—***67.19***—					
LLaMA 7B	48.84	**49.47**	50.53	**53.85**	48.66	**52.17**
LLaMA 65B	**60.22**	59.77	60.18	**64.94**	63.47	**65.33**
BLOOM 176B	42.02	**43.17**	46.33	**48.09**	45.99	**50.00**
BLOOMZ 176B	49.31	**49.60**	45.91	**50.73**	49.22	**50.53**

(b) English-Italian

Table 4: 1-shot, 3-shot and 5-shot results for En-Es and En-It prompting with randomised examples (`Rand.`) versus similar contexts (`Sim.`). The best-performing systems from Table 3, i.e. DeepL and NLLB-200 are chosen as baselines. For LLMs, for each setting, the better-performing baseline between `Rand.` and `Sim.` is highlighted in bold. The overall best score (among all LLMs) is underlined as well, while the best NMT system is also italicized.

4.4.2 Fine-tuning with ambiguous corpora

We fine-tune Alpaca 7B, BLOOM 7B and BLOOMZ 7B in En-Es and En-It directions using the data described in Section 4.2. We show our results when prompting these fine-tuned LLMs in Table 5. We make the following observations:

1. **Fine-tuning generally improves performance.** We observe that fine-tuned LLMs significantly outperform their non-finetuned versions in most cases. The biggest improvement is observed for BLOOM 7B in En-It, where accuracy increases by as high as 47.73%, indicating the effectiveness of our method. The only exception to this is when the LLM is already strong, such as BLOOMZ 7B at En-Es, and then the improvements are marginal. But even so, strong instruction-tuned LLMs like BLOOMZ still gain significantly from fine-tuning on the En-It pair – where it was originally weaker due to Italian being an unseen language during pretraining.

2. **Best Cross Entropy does not necessarily translate to best disambiguation accuracy.** Looking at Table 5, we note that the checkpoints with the best cross-entropy fall short of the topline with the best DiBiMT accuracies, suggesting the former is not an optimal metric for this task. Future work could benefit from using disambiguation-specific metrics for validation, leveraging other ambiguous test sets like MuCoW (Raganato et al., 2020b).

3. **Fine-tuning for 2-3 epochs is sufficient.** We plot the DiBiMT accuracy versus epoch curves in Figure 2 where the performance is evaluated after each epoch. We observe that in all cases, accuracy peaks between the 1st and the 3rd epoch, after which it mostly plateaus or dips slightly - suggesting that one does not need to fine-tune these LLMs for too long.

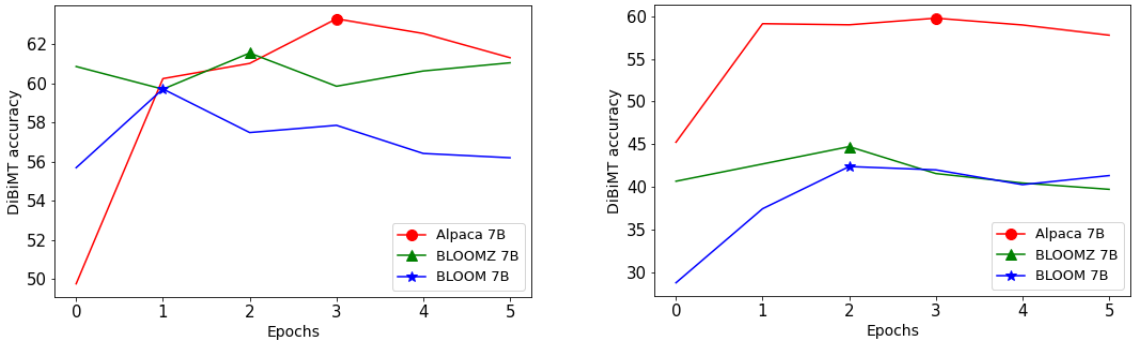
4. **Fine-tuning improves LLM performance until about 65K training samples.** We now try to answer the Research Question of how many training samples we need for fine-tuning these LLMs, to get optimal performance. We plot the Accuracy vs corpus size graph in Figure 3, where we indicate corpus size by the number of parallel sentences. We observe that accuracy increases non-monotonically with an increase in corpus size, but peaks anywhere between 36K-63K training samples, which seems to depend on the pre-existing capabilities of the LLM. For a raw foundation LLM like BLOOM 7B, relatively more fine-tuning data (54K-63K) appears to be beneficial. Alpaca 7B, which has been instruction-tuned on an English-only dataset, also seems to benefit from further fine-tuning—especially for En-Es, accuracy peaks after 63K training samples. However, for a powerful LLM like BLOOMZ that has been instruction-tuned on a large multilingual dataset like xP3 (Muennighoff et al., 2023), fine-tuning on smaller datasets (at most 36K sentences, in our case) appears to suffice.

4.5 Overall MT performance of disambiguation-adapted LLMs

Lastly, for completeness, we evaluate the overall translation quality of the key LLMs used in this work, since we are interested in noting how well the reported disambiguation accuracies extend to overall MT performance. For our test set, we want to choose one recently released (ideally within the

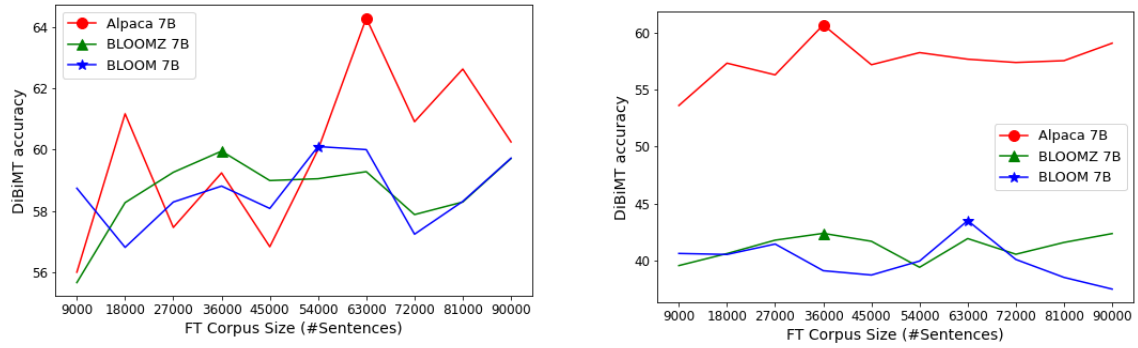
System	En-Es			En-It		
	Alpaca 7B	BLOOM 7B	BLOOMZ 7B	Alpaca 7B	BLOOM 7B	BLOOMZ 7B
w/o FT	49.75	55.69	60.87	45.24	28.79	40.68
FT (Best Cross-Entropy Loss)	63.27	57.86	60.39	59.62	37.72	39.73
FT (Best Attained Acc.)	63.31	59.72	61.56	59.77	42.40	44.73

Table 5: DiBiMT Accuracies after fine-tuning Alpaca 7B, BLOOM 7B, and BLOOMZ 7B on En-Es and En-It pairs. The second row indicates checkpoints with the best cross-entropy loss on the validation set, while the last row shows the one with the best attained DiBiMT accuracy when evaluating after each epoch, and serves as a "topline".



(a) English-Spanish (b) English-Italian

Figure 2: DiBiMT accuracy at the end of every epoch, for the LoRA fine-tuned LLMs



(a) English-Spanish (b) English-Italian

Figure 3: DiBiMT accuracy vs fine-tuning (FT) corpus size in terms of parallel sentence count. These results are obtained from evaluating checkpoints at every 300 steps in the 1st epoch - which roughly corresponds to about 9K sentences, since we use a batch size of 32.

last year) to minimize the chances of its inclusion in the pretraining corpora of LLMs. We, thus, use FLORES 200 (NLLB Team et al., 2022) as our test set since it satisfies this criterion and also supports all our languages of evaluation. We use spBLEU[12] (Goyal et al., 2022), chrF++[13] (Popović, 2017) and COMET22 (Rei et al., 2022) using the `wmt22-comet-da` model as metrics. In this setting, we evaluate Alpaca with 0-shot prompting, while LLaMA 7B, LLaMA 65B and BLOOM 176B use the 1-shot setup. NLLB-200 is our primary supervised NMT baseline. We also evaluate LoRA fine-tuned versions of Alpaca 7B and BLOOM 7B, from section 4.4.2, on the English-Spanish and English-Italian pairs. We exclude BLOOMZ from this evaluation since it is instruction-tuned on FLORES200. We report our results in Table 6.

We observe trends similar to those of our DiBiMT experiments. BLOOM 176B performs well in translation of seen languages, performing comparably to NLLB-200 in English-Spanish and outperforming it in English-Chinese. This is particularly the case for COMET22 scores, a metric which has shown high correlations with human

[12] `nrefs:1|case:mixed|eff:no|tok:flores101|smooth:exp|version:2.3.1`

[13] `nrefs:1|case:mixed|eff:yes|nc:6|nw:2|space:no|version:2.3.1`

System	En-Es			En-It		
	spBLEU	chrF++	COMET22	spBLEU	chrF++	COMET22
NLLB-200 54B	32.50	53.79	0.86	37.60	57.33	0.89
Alpaca 7B (0-shot)	23.90	47.30	0.83	23.30	46.40	0.83
LLaMA 7B (1-shot)	23.20	46.20	0.82	22.10	45.00	0.82
LLaMA 65B (1-shot)	27.20	49.70	0.83	28.50	50.50	0.85
BLOOM 7B (1-shot)	24.00	46.30	0.82	10.00†	33.40†	0.63†
BLOOM 176B (1-shot)	28.60	51.20	0.85	20.80†	45.20†	0.81†
Alpaca 7B (FT, 0-shot)	27.40	50.20	0.85	29.20	51.40	0.87
BLOOM 7B (FT, 0-shot)	28.70	51.00	0.86	20.90	45.80	0.80

System	En-Zh			En-Ru			En-De		
	spBLEU	chrF++	COMET22	spBLEU	chrF++	COMET22	spBLEU	chrF++	COMET22
NLLB-200 54B	23.10	22.83	0.82	38.00	56.34	0.90	44.80	62.79	0.88
Alpaca 7B (0-shot)	4.80†	10.40†	0.62†	21.80	42.60	0.82	27.30	50.30	0.82
LLaMA 7B (1-shot)	5.60†	10.80†	0.66†	20.70	41.20	0.79	22.80	45.40	0.78
LLaMA 65B (1-shot)	13.80†	17.60†	0.77†	26.70	46.10	0.82	31.80	52.80	0.81
BLOOM 7B (1-shot)	19.00	19.50	0.83	3.70†	22.30†	0.46†	8.20†	31.70†	0.51†
BLOOM 176B (1-shot)	25.10	23.80	0.86	10.30†	31.80†	0.65†	19.90†	45.40†	0.74†

Table 6: FLORES 200 results for k-shot prompting of some key LLMs used in this work, compared with the NLLB-200 baseline. We also include results for the LoRA fine-tuned models, for the En-Es and En-It pairs. Same as the previous notation, we indicate all unseen language results with a †. We observe similar trends in all standard MT metrics, as those observed with DiBiMT accuracy.

	spBLEU w/ acc.	ChrF++ w/ acc.	COMET22 w/ acc.
ρ	0.83	0.56	0.76
p-value	0.0001	0.0039	0.0010

Table 7: Pearson's correlation ρ (Benesty et al., 2009) between DiBiMT accuracy and spBLEU, chrF++, and COMET22 respectively, together with p-values.

evaluation, ranking second in the WMT22 Metrics shared task (Freitag et al., 2022). For the other languages, LLaMA 65B usually performs better than BLOOMZ, but in the 1-shot prompting setup, it is unable to beat the NLLB-200 54B MOE. We also notice that the fine-tuned versions of Alpaca 7B and BLOOM 7B consistently outperform their vanilla counterparts – suggesting our techniques to improve disambiguation performance also boost overall translation quality.

Thus, while we evaluate key LLMs to verify consistency in trends, we avoid re-running all our baselines on FLORES200. Instead, we try to answer a broader question: how well does DiBiMT disambiguation accuracy correlate with standard MT metrics? We conduct a Pearson's correlation test (Benesty et al., 2009) between the accuracy metric and spBLEU, chrF++, and COMET22 respectively. We report our results in Table 7, and find that all MT quality metrics correlate positively with accuracy—with p-values of the two-sided alternative hypothesis being much lesser than 0.05 in all cases. We discover that spBLEU and COMET22 exhibit higher correlations than chrF++. We hypothesize that this could be due to the character-level chrF++ being less sensitive to word-level senses. Overall, the results of Tables 6 and 7 suggest that the significant accuracy improvements noted earlier are not at the cost of translation quality, and in turn, could yield improvements in overall MT scores too.

5 Conclusion

In this work, we studied the capabilities of LLMs to handle ambiguity during machine translation. We choose seven of the most widely used foundation and instruction-tuned LLMs and compare accuracy with SOTA commercial and open-source NMT systems on the DiBiMT translation benchmark. Out of 5 language directions, we report scores comparable to the SOTA on two (En-Ru, En-It) and set a new SOTA on two others (En-Zh, En-Es). We then present two techniques that significantly improve disambiguation accuracy: in-context learning with similar contexts, and fine-tuning on an ambiguous corpus. We end the paper with an evaluation of overall MT quality. We hope the methods and findings shared in this work could guide future researchers studying ambiguity in translation.

Limitations

In this work, we attempt to note overall trends in LLM performance as compared to conventional NMT systems and, based on our results, suggest methods that generally improve performance. That said, there are exceptions to these trends - prompting with similar contexts can, at times, degrade performance and so can increasing the number of demonstrations (see Table 4). But there is some consistency here too that these observations mostly apply to smaller LLMs (such as LLaMA 7B) while the larger LLMs benefit more significantly. Also, as noted in Section 4.4.1, in a small percentage of cases (7.5%), we are unable to find 5 matches when attempting 5-shot prompting with similar contexts. In such cases, it might be worthwhile, from a performance perspective, to use random demonstrations; nonetheless, since we are interested in verifying the utility of similar contexts and also since there are only a few cases where it might be pertinent, we do not explore this.

Acknowledgements

This work has received funding from UK Research and Innovation under the UK government's Horizon Europe funding guarantee [grant numbers 10039436 and 10052546].

The computations described in this research were performed using the Baskerville Tier 2 HPC service (https://www.baskerville.ac.uk/). Baskerville was funded by the EPSRC and UKRI through the World Class Labs scheme (EP/T022221/1) and the Digital Research Infrastructure programme (EP/W032244/1) and is operated by Advanced Research Computing at the University of Birmingham.

References

Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 2021. ESC: Redesigning WSD with extractive sense comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4661–4672.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 37–40. Springer.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Niccolò Campolungo, Federico Martelli, Francesco Saina, and Roberto Navigli. 2022. DiBiMT: A novel benchmark for measuring Word Sense Disambiguation biases in Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4331–4352, Dublin, Ireland. Association for Computational Linguistics.

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 61–72, Prague, Czech Republic. Association for Computational Linguistics.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 33–40.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2023. Iterative translation refinement with large language models. *arXiv preprint*.

Heeyoul Choi, Kyunghyun Cho, and Yoshua Bengio. 2017. Context-dependent word representation for neural machine translation. *Computer Speech & Language*, 45:149–160.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. PaLM: Scaling language modeling with pathways. *arXiv preprint*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1):4839–4886.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023. Exploring human-like translation strategy with large language models. *arXiv preprint*.

Amr Hendy, Mohamed Gomaa Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? a comprehensive evaluation. *arXiv preprint*.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *arXiv preprint*.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Frederick Liu, Han Lu, and Graham Neubig. 2018. Handling homographs in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1336–1345, New Orleans, Louisiana. Association for Computational Linguistics.

Yasmin Moslem, Rejwanul Haque, and Andy Way. 2023. Adaptive machine translation with large language models. *EAMT*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Roberto Navigli, Michele Bevilacqua, Simone Conia, Dario Montagnini, and Francesco Cecconi. 2021. Ten years of BabelNet: A survey. In *IJCAI*, pages 4559–4567.

NLLB Team, Marta Ruiz Costa-jussà, James Cross, Onur cCelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Alison Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon L. Spruit, C. Tran, Pierre Yves Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzm'an, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *ArXiv*, abs/2207.04672.

Jonathan Pilault, Xavier Garcia, Arthur Bražinskas, and Orhan Firat. 2023. Interactive-chain-prompting: Ambiguity resolution for crosslingual conditional generation with interaction. *arXiv preprint*.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Xiao Pu, Nikolaos Pappas, James Henderson, and Andrei Popescu-Belis. 2018. Integrating weakly supervised word sense disambiguation into neural machine translation. *Transactions of the Association for Computational Linguistics*, 6:635–649.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2019. The mucow test suite at wmt 2019: Automatically harvested multilingual contrastive word sense disambiguation test sets for machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 470–480.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2020a. An evaluation benchmark for testing the word sense disambiguation capabilities of machine translation systems. In *Proceedings of The 12th Language Resources and Evaluation Conference*. European Language Resources Association (ELRA).

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2020b. An evaluation benchmark for testing the word sense disambiguation capabilities of machine translation systems. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3668–3675, Marseille, France. European Language Resources Association.

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan. 2023. Do GPTs produce less literal translations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1041–1050, Toronto, Canada. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint*.

Yves Scherrer, Alessandro Raganato, and Jörg Tiedemann. 2020. The MUCOW word sense disambiguation test suite at WMT 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 365–370, Online. Association for Computational Linguistics.

Gongbo Tang, Rico Sennrich, and Joakim Nivre. 2019. Encoders help you disambiguate word senses in neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1429–1435, Hong Kong, China. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An instruction-following LLaMA model. `https://github.com/tatsu-lab/stanford_alpaca`.

Jörg Tiedemann and Santhosh Thottingal. 2020. Opus-mt–building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Warren Weaver. 1952. Translation. In *Proceedings of the Conference on Mechanical Translation*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR.

Source	Tap the head of the drum for this roll.	
DeepL	敲击鼓的头部进行滚奏。	*head, literal*
BLOOMZ	敲击鼓面，发出这个鼓点。	*striking surface of a drum*
Source	they tracked him back toward the head of the stream.	
DeepL	他们跟踪他回到溪头.	*head (literal and unnatural)*
BLOOMZ	他们跟着他回到了河的上游。	*upstream*
Source	The expedition followed the river all the way to the head.	
DeepL	探险队沿着河流一直走到河头。	*head (literal and unnatural)*
BLOOMZ	探险队顺着河一直走到源头。	*source*
Source	How much head do you have at the Glens Falls feeder dam?	
DeepL	格伦瀑布支坝的水头有多大?	*hydraulic head*
BLOOMZ	你有多少头牛在格伦瀑布的蓄水池里?	*(a classifier word to express quantities (of cows))*
Source	the office was full of secret heads.	
DeepL	办公室里到处都是秘密头目.	*leader, ringleader*
BLOOMZ	办公室挤满了神秘的首脑。	*leader, head of state*

Table 8: Manual inspection on English-to-Chinese translation focused on the disambiguation of "head", corresponding to the first five test instances in DiBiMT. The baselines are DeepL and BLOOMZ 176B, the highest performing NMT system and LLM for this pair (from Table 3). The reported annotations are obtained from a native Chinese speaker who was invited to label the sense of the translated ambiguous word.

A Appendix

A.1 Qualitative comparison: BLOOMZ vs DeepL

We choose the best-performing LLM and the SOTA MT system from Table 3 – focusing on the En-Zh pair since LLMs seem to yield the highest gains there. With the help of a native Chinese speaker, we got hypotheses from these two systems annotated, for the first 5 sentences of the DiBiMT test set. We observe that although there are cases where DeepL gets it right over BLOOMZ (example 4) or where both are correct (Example 5), in many instances BLOOMZ appears to generate more contextual (and less literal) translations. We hypothesize that this could potentially be due to the former's powerful language modelling abilities

A.2 Trade-off between prompting and fine-tuning

We show in Section 4.4 that both prompting with similar contexts through In-Context Learning (ICL) and LoRA fine-tuning can significantly improve performance. However, depending on the use case, it might be better to favour one over the other. For instance, in production environments, LLMs that are LoRA fine-tuned on ambiguous text can provide powerful disambiguation performance, while also being more feasible to deploy and run at scale. In contrast, ICL with k-shot prompting, especially for higher values of k, can significantly increase query size and memory consumption, necessitating reduced batch size and thus, throughput.

However, conducting ICL with similar ambiguous contexts can be used to query LLMs as large as LLaMA 65B and BLOOMZ 176B and yield performance comparable to SOTA MT systems (see Table 4). The preprocessing cost overhead of such a method, namely disambiguating the test set, is also low - it took us about 13 seconds to disambiguate a test set of about 500 sentences on 1 Nvidia GeForce RTX 3090. In contrast, the one-time cost of fine-tuning can be quite expensive—for instance, it took us 44 hours to fine-tune an Alpaca 7B with LoRA on a single Nvidia Tesla A100 40G. Thus, in GPU-scarce settings where the costs of LoRA fine-tuning are prohibitive, it might be favourable to use ICL to query massive LLMs and obtain SOTA performances. In contrast, production environments are likely to prefer the fine-tuned LLMs, since the one-off fine-tuning costs can be amortized.

A Closer Look at Transformer Attention for Multilingual Translation

Jingyi Zhang[1], **Hongfei Xu**[2], **Kehai Chen**[3] **and Gerard de Melo**[1]
[1]Hasso Plattner Institute, University of Potsdam, Germany
[2]Zhengzhou University, Henan, China
[3]School of Computer Science and Technology, Harbin Institute of Technolgy, Shenzhen, China
Jingyi.Zhang@hpi.de, hfxunlp@foxmail.com, chenkehai@hit.edu.cn
gerard.demelo@hpi.de

Abstract

Transformers are the predominant model for machine translation. Recent studies also showed that a single Transformer model can be trained to learn translation for multiple different language pairs, achieving promising results. In this work, we investigate how multilingual Transformer models pay attention when translating different language pairs. To achieve this, we first conduct automatic pruning to eliminate a large number of noisy heads and then assess the functions and behaviors of the remaining heads in both self-attention and cross-attention. We find that different language pairs, in spite of having different syntax and word orders, tend to share the same heads for the same functions, such as syntax heads and reordering heads. However, the different characteristics of different language pairs can clearly cause interference in function heads and affect head accuracies. Additionally, we reveal an interesting behavior of the Transformer cross-attention: the deep-layer cross-attention heads work in a cooperative way to learn different options for word reordering, which may be caused by the nature of translation tasks having multiple different gold translations in the target language for the same source sentence.[1]

1 Introduction

For traditional statistical machine translation, such as phrase-based translation (Koehn et al., 2003), the translation process is very clear: source phrases are translated into target phrases according to translation rules and then target phrases are reordered to ensure the fluency of the target sentence. However, in state-of-the-art neural translation models (Bahdanau et al., 2014; Vaswani et al., 2017; Chen et al., 2018), how the model learns to translate is substantially less obvious. The behavior of the Transformer model (Vaswani et al., 2017) remains particularly hazy, as it contains many different self- and cross-attention heads in different layers.

A number of existing studies conducted analyses of functions and behaviors of attention heads in Transformer translation models. Voita et al. (2019b) found that the Transformer attention is noisy, as most of the Transformer heads can be pruned away without significant loss in translation quality. They also identified three important functions of self-attention in the Transformer encoder, such as heads focusing on syntax. Ferrando and Costa-jussà (2021) demonstrated that the cross-attention of the Transformer model frequently attends to uninformative source words to balance the contribution of source and target context for predicting the next word. Chen et al. (2020) showed that some cross-attention heads learn alignment for the current target word, achieving higher accuracies than cross-attention heads that learn alignment for the next target word. However, these methods only analyzed attention in bilingual models, not for multilingual Transformer models.

Multilingual translation, i.e., training a single Transformer to learn translation for multiple different language pairs, has received much attention in recent years and obtained promising results (Wang et al., 2020; Kim et al., 2021; Pires et al., 2023). A number of studies investigated how a multilingual Transformer learns to translate different language pairs. Several of these (Lin et al., 2021; Wang et al., 2020; Xie et al., 2021) learned language-dependent weight masks to identify language-dependent subnetworks. Pires et al. (2023) trained the multilingual Transformer to learn language-specific layers and improved translation quality. Chiang et al. (2022) and Kim et al. (2021) assessed how different language pairs share important heads in multilingual Transformer models.

However, prior work has not yet studied the specific functions and behaviors of different attention heads in multilingual Transformer models. In this

[1]Code and scripts for reproducing our results can be found https://github.com/jingyiz/multilingual-translation-attention-head-analysis.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 496–506
December 6–7, 2023.

paper, we investigate functions and behaviors[2] of both self-attention and cross-attention for multilingual translation. We find that different language pairs with different syntax and different word orders tend to share the same heads for the same functions (such as syntax heads and reordering heads), but the different characteristics of different language pairs can clearly cause interference in function heads and affect head accuracies compared to bilingual models. We further obtain an interesting finding about how the Transformer learns word reordering: different cross-attention heads in deep layers work in a cooperative way to learn different options for reordering. This may result from the fact that there are multiple different gold translations (reorderings) in the target language for the same source sentence[3].

2 Related Work

There are a number of studies on analyzing layer representations of different Transformer layers. Voita et al. (2019a) used canonical correlation analysis and mutual information estimators to study how information flows across Transformer layers for different learning objectives. Kudugunta et al. (2019) used Singular Value Canonical Correlation Analysis (SVCCA) to analyze how representations evolve in a multilingual translation model. Xu et al. (2021b) analyzed how word translation evolves in Transformer layers and showed that translation already happens progressively in encoder layers and even in the input embeddings, by measuring word translation accuracy of different Transformer layers. These methods did not analyze the specific functions of attention heads.

Other prior work analyzed Transformer attention to better understand a particular aspect of the translation process. Tang et al. (2021) analyzed Transformer attention for negation translation and showed that negation is often rephrased during training, which can make it more difficult for the model to learn a reliable link between source and target negation. Tang et al. (2018) analyzed Transformer cross-attention for learning word sense disambiguation (WSD) and showed that cross-attention is likely to distribute more attention to the ambiguous noun itself rather than context tokens, in comparison to other nouns, which suggests that the Transformer learns to encode contextual information necessary for WSD in the encoder hidden states. Additionally, Tang et al. (2018) also noticed that, from shallow layers to deep layers, the cross-attention accuracy for aligning the next target word first increases and then decreases. However, we our study is the first to reveal the cooperative behavior of cross-attention heads.

There is also prior work that studied representation sharing in multilingual translation. Firat et al. (2016) proposed a multiway, multilingual model with language-specific encoders and decoders and showed result quality improvements over models trained on only one language pair. Several authors (Zhang et al., 2021; Bapna and Firat, 2019; Zhu et al., 2021) considered language-dependent gating and adaptation for layer representations. Xu et al. (2021a) proposed parallel encoder and decoder layers with language-dependent weighted layer aggregation. Wang et al. (2019) presented a universal representer to replace both encoder and decoder models to enable parameter sharing between encoder and decoder and they made the representer sensitive for specific languages using language-sensitive embedding, attention, and discriminator. Zhu et al. (2020) incorporated a language-aware interlingua into the encoder–decoder architecture, which enables the model to learn a language-independent representation from the semantic spaces of different languages, while still allowing for language-specific specialization of a particular language pair. Additionally, Shaham et al. (2023) showed that controlling the proportion of each language pair in the training data can balance the amount of interference between languages in multilingual models. Yuan et al. (2023) developed a detachable model by assigning each language (or group of languages) to an individual branch that supports plug-and-play training and inference with a novel efficient training recipe. Xu et al. (2023) investigated how to utilize intra-distillation to learn more language-specific parameters and then showed the importance of these language-specific parameters. However, these methods did not investigate the head functions in multilingual models.

[2]Following Voita et al. (2019b)'s work, we use a weight-based method for analyzing attention head behaviors. It is also possible to use a norm-based method (Kobayashi et al., 2020), which may provide a more detailed interpretation of the inner workings of Transformers compared to weight-based methods in some cases.

[3]In the training data of translation models, it is rather rare that the same source sentence has multiple different translated target sentences, but it is very common that the same source phrase has multiple different translated target phrases. Therefore, translation models are able to learn to translate a source sentence into different target sentences.

	DeEn	FrEn	RoEn	EnDe	EnFr	EnRo	Average
R-bi	25.90	29.47	31.46	21.94	31.08	25.74	27.59
R-multi	25.88	29.85	34.07	21.70	31.00	26.17	28.11
R-finetune	26.48	29.91	34.96	22.61	31.79	27.21	28.82
R-prune ($\lambda = 25$)	26.31	29.76	34.87	22.31	31.46	27.12	28.63
R-prune ($\lambda = 35$)	26.23	29.56	34.82	22.05	31.40	27.21	28.54

Table 1: Translation results (BLEU) on the test sets.

Train		
DeEn	2.5M	Europarl v7, TED2020, News-Commentary v11
FrEn	2.5M	Europarl v7, TED2020, News-Commentary v11
RoEn	0.9M	Europarl v8, TED2020, SETIMES2
Valid		
DeEn	5,014	newstest2009, newstest2010
FrEn	5,014	newstest2009, newstest2010
RoEn	1,999	newsdev2016
Test		
DeEn	9,006	newstest2011, newstest2012, newsdev2013
FrEn	9,006	newstest2011, newstest2012, newsdev2013
RoEn	1,999	newstest2016

Table 2: Datasets and their number of sentence pairs.

	$\lambda = 25$	$\lambda = 35$
DeEn	74	54
FrEn	58	49
RoEn	74	52
EnDe	85	64
EnFr	65	53
EnRo	81	56
Shared	55	45
Total	144	

Table 3: Number of remaining heads after automatic pruning for different translation directions. "Shared" means the number of heads that remain for all six translation directions. "Total" refers to the original number of all heads before automatic pruning.

3 How Do Transformers Pay Attention for Multilingual Translation?

3.1 Methodology and Experimental Setup

Bilingual Baseline. We used the original Transformer model in its base setting (Vaswani et al., 2017) (i.e., the same model parameters, training parameters and inference parameters) as our bilingual baseline model and conducted translation experiments for six translation directions[4]: German↔English (De↔En), French↔English (Fr↔En), and Romanian↔English (Ro↔En). For each translation direction (such as De→En), we trained a Transformer model using the training data and validation data for this translation direction as shown in Table 2.[5] Following Vaswani et al. (2017), we trained each model for 100k training steps. However, because our training data size for a single translation direction is smaller than in their work, 100k training steps caused overfitting in our models. Therefore we computed the validation loss after each training epoch and then chose the best validation checkpoint for evaluation. Translation results on the test sets are given in Table 1 as R-bi.

Multilingual Translation. For multilingual translation, we trained a single Transformer to learn translation for all six translation directions. We combined all training data in Table 2 together and added a special token at the beginning of each source sentence to indicate which target language we desire the model to generate, following Johnson et al. (2017). We used the same base setting of the original Transformer with 100k training steps for our multilingual model. During training of the multilingual model, we computed the validation loss for the combined validation data after each training epoch and found that the validation loss continuously decreased, so we used the final checkpoint of the multilingual model for evaluation. The evaluation results of the multilingual model are given in Table 1 as R-multi. In the results, we can observe that the multilingual model obtained comparable or higher translation quality compared to our bilingual baseline for different language pairs.

Finetuning. We then finetuned[6] the multilingual model for each translation direction using direction-

[4]We chose these language pairs because parallel sentences with gold-standard word alignments are available (Zhang and van Genabith, 2021) for these language pairs, which can be used to analyze target-to-source attention (alignment).

[5]For subword segmentation, we applied byte pair encoding (Sennrich et al., 2016) and learned a joint vocabulary of size 32k for all languages in our experiments.

[6]Finetuning a multilingual model for a given translation direction (i.e., multilingual pretraining) is very popular for low-resource language pairs and can significantly improve translation quality. We find that finetuning generally did not change the functions of different heads (see Figure 1) but did improve the accuracies of function heads for the given translation direction (see Table 12).

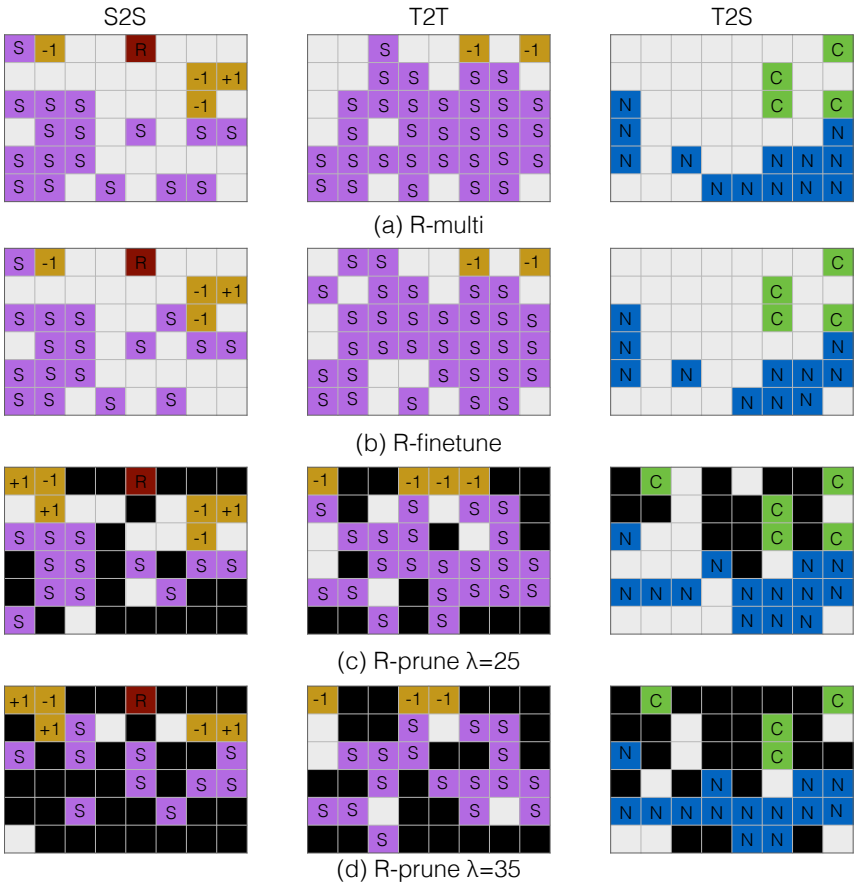


Figure 1: Functional heads (red: rare word head (R); yellow: positional head (-1 and $+1$); purple: syntactical head (S); green: C-alignment head (C); blue: N-alignment head (N)) contained in (a) the multilingual model (R-multi); (b) finetuned models (R-finetune); (c) pruned models with $\lambda = 25$; (d) pruned models with $\lambda = 35$. From left to right, the three columns of figures represent S2S, T2T, and T2S attention. Each figure shows attention heads from the first layer to the last layer (top-down) and each layer contains 8 heads. Black denotes heads that are pruned away.

specific training and validation data.[7] During finetuning, we set the maximum number of finetuning steps to 50k and computed the validation loss after each training epoch, and finally used the best validation checkpoint for evaluation. We found that models for all six directions converged during finetuning (the best validation checkpoint is not the last checkpoint). The results of the finetuned models are given in Table 1 as R-finetune. As shown in Table 1, finetuning a pre-trained multilingual translation model for a specific translation direction can improve the translation quality for the given translation direction (i.e., R-finetune $>$ R-multi). Table 1 also shows that the finetuned models can achieve higher translation quality compared to the bilingual models (R-bi) for all translation directions in our experiments.

Head Pruning. As shown by Voita et al. (2019b), Transformer attention is noisy, i.e., many attention heads carry no important function and can be pruned away without significant loss in translation quality. Following them, we conduct automatic pruning to identify important heads and analyze their functions. For each translation direction, we continue to finetune the already converged model (R-finetune) with a regularization loss (Louizos et al., 2017) along with the original translation loss to prune away useless heads. With the regularization loss, the model learns a 0/1 gate for each head. Heads with a 0 gate are pruned away. A weight λ is assigned to the regularization loss to control the amount of heads to be pruned, i.e., a higher weight for the regularization loss will result in more heads being pruned away. Translation results after head pruning are given in Table 1 and the number of remaining heads for each translation direction is

[7] For example, when we finetuned the multilingual model for the De→En direction, we only used training and validation data with German in the source and English in the target.

	DeEn	FrEn	RoEn	EnDe	EnFr	EnRo
Accu	0.31	0.36	0.19	0.33	0.36	0.26

Table 4: Accuracy of the rare word head.

listed in Table 3.[8] At $\lambda = 35$, roughly $2/3$ of all heads were pruned away and the average BLEU only decreased by 0.28. Table 3 also shows that different language pairs tended to share important heads, as most of the remaining heads remained for all six translation directions.

3.2 Head Function Analysis

We analyzed the behavior of the remaining heads to understand their functions.

Source-to-source Rare-word Heads. We find that one source-to-source (S2S) attention head in the first encoder layer tends to attend to the most infrequent word of the input sentence, which agrees with the bidirectional findings of Voita et al. (2019b). The maximum weight of this head is assigned to one of the least two frequent words in the input sentence roughly 30% of the time, as shown in Table 4 for most language pairs. We also find that this behavior of attending to rare words does not occur in target-to-target (T2T) and target-to-source (T2S) attention, as all T2T and T2S heads achieved less than 10% accuracy at attending to the two least frequent words. The S2S rare word head is marked in red in Figure 1.[9]

Self-attention Positional Heads. We find that some self-attention heads in both the encoder and the decoder tend to attend to neighbors ($+1$ or -1 position). We call a self-attention head "positional" if its maximum attention weight is assigned to neighbors at least 80% of the time. For example, if the maximum weight of a head is assigned to the -1 relative position more than 80% of the time, then this head is identified as a positional -1 head, as shown in Figure 1. Table 5 shows positional heads found in the finetuned models (R-finetune). We find that different language pairs generally share the same positional heads, and positional heads only occur in shallow encoder and decoder layers. As shown in Figure 1, positional heads essentially remain unchanged during finetuning. However, during pruning, some positional heads are eliminated and some new positional heads emerge, mostly because positional attention is easy to learn and therefore this function tends to migrate from one head to another during automatic pruning.

	head	directions
S2S (-1)	1:6	DeEn,FrEn,RoEn,EnDe,EnFr,EnRo
	2:6	DeEn,FrEn,RoEn,EnDe,EnFr,EnRo
	0:1	EnDe,EnFr
S2S ($+1$)	1:7	DeEn,FrEn,RoEn,EnDe,EnFr,EnRo
T2T (-1)	0:5	DeEn,FrEn,RoEn,EnDe,EnFr,EnRo
	0:7	DeEn,FrEn,RoEn,EnDe,EnFr,EnRo

Table 5: Positional heads in the finetuned models (R-finetune). "1:6" denotes the 6th head in the 1st layer.

	German	French	Romanian	English
obj	1	-2	-1	-2
nsubj	1	1	2	1
advmod	1	1	1	1
amod	1	-1	-1	1

Table 6: The highest-probability relative distance for different dependency relationships (forward direction).

Self-attention Syntactical Heads. We find that some self-attention heads in both the encoder and the decoder learn syntactical dependencies, i.e., the maximum attention weight is assigned to a syntactically related word of the current word. We call a self-attention head "syntactical" if it learns a dependency relationship with an accuracy at least 10% higher than the baseline accuracy of this relationship. The baseline accuracy of a dependency relationship is the accuracy of a fictional head that always attends to the most likely relative position of this relationship. For example, for the obj dependency relationship in English, the correct dependency typically is encountered at the -2 relative position (38% of cases), which is the most likely relative position for this relationship. Hence, a fictional head that always attends to the -2 relative position will achieve 38% accuracy for this relationship, and 38% can serve as the baseline accuracy for the English obj relationship. For different languages, the most likely relative position of the obj relationship is different, as shown in Table 6. We look at four important dependency relation-

[8] The base Transformer model contains 144 attention heads in total: 48 self-attention heads in the encoder, 48 self-attention heads and 48 cross-attention heads in the decoder. Cross-attention and self-attention in both the encoder and the decoder have 6 layers and each layer contains 8 heads.

[9] Figure 1 shows functional heads identified for at least one translation direction. For example, all syntactical heads identified for different translation directions as shown in Table 7 and Table 9 are marked as S heads in Figure 1. Black heads are heads that were pruned away for all translation directions during automatic pruning.

✓	3:6	**amod-f**	DeEn,FrEn
		amod-b	FrEn,RoEn
		advmod-f	FrEn
		nsubj-f	EnDe
×	3:2	**obj-b**	DeEn,RoEn,EnDe,EnFr
		nsubj-f	DeEn
✓	2:0	**obj-b**	RoEn,EnDe,EnFr
		nsubj-f	DeEn,EnDe
✓	2:2	**obj-f**	RoEn,EnDe,EnFr,EnRo
×	5:5	**obj-b**	DeEn,RoEn,EnRo
		nsubj-f	DeEn
✓	4:2	**obj-f**	DeEn,RoEn,EnDe
×	2:1	**amod-f**	DeEn
		amod-b	RoEn
		obj-b	RoEn
×	4:1	**nsubj-f**	DeEn,EnDe,EnFr
✓	3:4	**obj-f**	RoEn,EnDe
		nsubj-b	DeEn
×	5:1	**obj-f**	RoEn
		advmod-b	RoEn
✓	5:0	**amod-b**	FrEn,RoEn
✓	0:0	**nsubj-f**	FrEn,EnDe
✓	3:7	**nsubj-b**	FrEn
×	2:5	**advmod-f**	FrEn
×	5:3	**obj-f**	RoEn
×	4:0	**nsubj-b**	RoEn
×	3:1	**obj-b**	RoEn

Table 7: Dependency relationships learned by S2S syntactical heads in the finetuned models. × means the head is pruned away with automatic pruning at $\lambda = 35$, while ✓ means the head remains after pruning.

head	DeEn	FrEn	RoEn	EnDe	EnFr	EnRo
2:2	0.06	0.43	0.44	**0.65**	**0.51**	**0.54**
4:2	**0.46**	**0.44**	**0.48**	0.53	0.45	0.48

Table 8: Accuracy of S2S heads for the obj-f relationship in the finetuned models. Among all S2S heads, head "2:2" achieved the highest obj-f accuracy for EnDe, EnFr, and EnRo; head "4:2" achieved the highest obj-f accuracy for DeEn, FrEn, and RoEn.

ships[10]: obj (v→o), nsubj (v→s), advmod (v→a), and amod (n→a). For each of these 4 relationships, we consider both the forward and the backward directions. Ultimately, we thus investigate whether a head learns any of the 8 relationships obj-f, obj-b, nsubj-f, nsubj-b, advmod-f, advmod-b, amod-f, and amod-b. We find a head can learn different dependency relationships, as shown in Tables 7 and 9. The most important dependency relationship learned by the Transformer is obj, as more than half of all syntactical heads mainly learn the obj relationship. Tables 7 and 9 further show that some translation directions share some syntactical heads (e.g., DeEn, RoEn, and EnDe share the

[10]We used the parsing results by the Stanford parser (Manning et al., 2014) as the ground truth label in our experiments.

✓	3:6	**obj-f**	DeEn,FrEn,RoEn,EnDe,EnRo
		advmod-f	DeEn,EnDe,EnFr
		amod-f	EnRo
		amod-b	EnRo
		nsubj-f	EnDe
✓	2:6	**obj-f**	DeEn,FrEn,RoEn,EnDe,EnFr,EnRo
		nsubj-f	EnDe
		advmod-f	EnDe
✓	3:5	**obj-f**	DeEn,FrEn,RoEn,EnDe,EnFr,EnRo
		nsubj-f	EnDe
		advmod-f	EnDe
✓	2:3	**nsubj-b**	DeEn,FrEn,RoEn,EnDe,EnFr,EnRo
		amod-f	EnRo
×	3:3	**nsubj-b**	DeEn,FrEn,RoEn,EnDe,EnFr,EnRo
✓	2:2	**obj-f**	DeEn,EnDe,EnRo
		nsubj-f	EnDe
		advmod-f	EnDe
✓	1:5	**amod-f**	EnFr,EnRo
		amod-b	EnRo
		advmod-f	EnFr
		obj-f	EnRo
✓	1:6	**obj-f**	DeEn,RoEn,EnDe,EnFr,EnRo
×	1:2	**advmod-f**	DeEn,EnDe,EnFr
		nsubj-f	EnDe
		obj-f	EnDe
✓	2:1	**amod-f**	EnFr,EnRo
		nsubj-f	EnDe
		advmod-f	EnDe
		obj-f	EnDe
✓	3:4	**obj-f**	DeEn,EnDe,EnRo
		nsubj-b	EnRo
×	5:5	**obj-f**	DeEn,FrEn,RoEn,EnDe
×	2:5	**obj-f**	EnDe,EnRo
		advmod-b	EnRo
×	0:2	**obj-f**	EnDe,EnRo
✓	4:5	**obj-f**	EnDe,EnRo
×	5:0	**obj-f**	DeEn,FrEn
×	4:4	**obj-f**	EnRo
		amod-f	EnRo
✓	1:0	**obj-f**	EnRo
		amod-b	EnRo
×	2:4	**nsubj-b**	DeEn
		amod-b	EnRo
✓	1:3	**nsubj-f**	EnDe
		advmod-f	EnDe
✓	4:0	**obj-f**	EnDe
		nsubj-f	EnDe
✓	3:7	**obj-f**	EnRo
		amod-b	EnRo
✓	4:6	**obj-f**	EnRo
✓	4:7	**obj-f**	EnDe
×	5:1	**obj-f**	EnDe
✓	4:1	**obj-f**	EnRo
×	5:3	**obj-f**	EnRo
×	5:6	**nsubj-b**	EnFr
×	3:1	**obj-f**	EnRo
×	0:1	**obj-f**	EnDe
×	2:7	**obj-f**	EnDe
✓	3:2	**obj-f**	EnRo

Table 9: Dependency relationships learned by T2T syntactical heads in the finetuned models. × means the head is pruned away with automatic pruning at $\lambda = 35$, while ✓ means the head remains after pruning.

Head	DeEn	FrEn	RoEn	EnDe	EnFr	EnRo
0:7	**0.88**	0.81	**0.86**	**0.85**	0.77	0.72
1:5	0.81	0.76	0.82	0.77	0.73	0.77
2:5	**0.88**	**0.84**	0.85	0.81	0.75	0.68
2:7	0.81	0.73	0.84	0.77	**0.80**	**0.87**

Table 10: Accuracy of C-alignment heads.

S2S "4:2" head for the `obj-f` relationship in Table 7), but not all translation directions share all syntactical heads. For a more direct overview of how different translation directions share syntactical heads, Table 8 gives the accuracy of different S2S heads for the `obj-f` relationship, which clearly shows that the "2:2" head mainly learns `obj-f` for EnDe, EnFr, and EnRo, while the "4:2" head mainly learns `obj-f` for DeEn, FrEn, and RoEn. Meanwhile, the S2S "2:2" head acquires nearly 0 `obj-f` accuracy for the DeEn direction, although this head is the most accurate `obj-f` head for EnDe, EnFr, and EnRo.

Cross-attention C-alignment Heads. We find that some cross-attention heads in the shallow layers learn word alignment for the current target word, i.e., the maximum weight is assigned to the source word aligned to the current target word. If a head achieves more than 80% accuracy for aligning the current target word, we call such a head a "C-alignment" head. When we calculate the alignment accuracy[11], we only consider situations when the current target word is a content word[12], as function words generally do not have clear alignments between different languages. By attending to the contextualized representation of the source word aligned to the current target word, C-alignment heads can help to retrieve the full context of the current target word (both the left-side and right-side context), in contrast to target-to-target self-attention, which can only attend to the left-side context of the current target word. Table 10 gives the accuracy of C-alignment heads in the finetuned models (R-finetune), showing that C-alignment heads generally learn the current word alignment for all translation directions, while the highest-accuracy C-alignment head for different directions may differ.

[11]We use human-annotated word alignments (Zhang and van Genabith, 2021) as the ground truth label for computing word alignment accuracies.

[12]For each language, we judge whether a word is a function word or a content word using a list of stopwords from NLTK, https://www.nltk.org/

Head	DeEn	FrEn	RoEn	EnDe	EnFr	EnRo
2:0	0.70	0.67	0.66	0.62	0.72	0.65
3:0	0.72	0.72	0.69	0.65	0.68	0.56
3:7	0.73	0.72	0.72	0.74	**0.77**	0.77
4:0	0.78	0.80	0.75	0.76	0.76	0.71
4:2	0.78	0.81	0.78	0.79	**0.77**	**0.78**
4:5	0.66	0.70	0.66	0.64	0.62	0.60
4:6	0.67	0.71	0.68	0.64	0.70	0.59
4:7	**0.82**	**0.83**	**0.81**	**0.81**	0.75	0.75
5:4	0.69	0.74	0.70	0.65	0.72	0.63
5:5	0.63	0.70	0.67	0.67	0.66	0.59
5:6	0.66	0.71	0.66	0.68	0.61	0.65

Table 11: Accuracy of N-alignment heads.

Cross-attention N-alignment Heads. We further find that some cross-attention heads in the deep layers learn word alignment for the next target word, i.e., the maximum weight is assigned to the source word aligned to the next target word. As the next word is unknown at the current decoding step, N-alignment heads are rather learning to predict the next target word than just aligning the next target word. Therefore, the N-alignment accuracies are generally lower than the C-alignment accuracies. We identify "N-alignment" heads as heads that achieve more than 70% accuracy for aligning the next word. When we calculate the N-alignment accuracy, we again only consider situations when the next target word is a content word, as before for the C-alignment accuracy. Figure 1 shows that C-alignment heads occur in shallow layers and N-alignment heads occur in deep layers, which indicates that the Transformer decoder appears to first use C-alignment heads to obtain the context of the current target word, and then, based on the context of the current target word, predicts which word to generate next. Table 11 gives the accuracy of N-alignment heads in the finetuned models (R-finetune), which shows that different language pairs generally shared N-alignment heads, which is surprising since the purpose of N-alignment heads is learning word reordering and different language pairs should have different reordering rules. Table 11 also shows that the highest-accuracy N-alignment heads are most likely from the 4th layer, i.e., the N-alignment accuracy first increases and then decreases as the layer number increases.

3.3 Head Behavior Analysis

We provide a further analysis of head behavior by comparing head accuracies in different multilingual and bilingual models. Table 12 gives the highest

			DeEn		FrEn		RoEn		EnDe		EnFr		EnRo	
S2S	obj-f	R-bi	**0.58**		0.40$^{\times}$		0.46		0.57		**0.55**		0.53	
		R-multi	0.42		0.43$^{\times}$		0.45		0.60		0.49		0.46$^{\times}$	
		R-finetune	0.46	△	**0.44**$^{\times}$	△	**0.48**	△	**0.65**	△	0.51	△	**0.54**	△
	obj-b	R-bi	**0.46**		**0.41**$^{\times}$		**0.60**		**0.69**		0.51		**0.61**	
		R-multi	0.41		0.37$^{\times}$		0.53		0.57		0.55		0.46	
		R-finetune	0.44	△	0.38$^{\times}$	△	0.57	△	0.59	△	**0.56**	△	0.50	△
T2T	obj-f	R-bi	**0.84**		0.76		0.69		**0.74**		0.72		0.53	
		R-multi	0.79		0.76		0.68		0.63		**0.73**		0.52	
		R-finetune	0.81	△	**0.78**	△	**0.70**	△	0.65	△	0.72	▽	**0.55**	△
	obj-b	R-bi	–		–		–		0.37$^{\times}$		–		–	
		R-multi	–		–		–		**0.38**$^{\times}$		–		–	
		R-finetune	–		–		–		0.37$^{\times}$	▽	–		–	
T2S	C-a	R-bi	**0.89**		**0.88**		**0.89**		**0.90**		**0.82**		**0.87**	
		R-multi	0.88		0.83		0.85		0.84		0.79$^{\times}$		0.86	
		R-finetune	0.88		0.84	△	0.86	△	0.85	△	0.80	△	**0.87**	△
	N-a	R-bi	0.80		**0.83**		0.79		0.81		**0.79**		**0.78**	
		R-multi	**0.83**		**0.83**		**0.82**		**0.83**		**0.79**		**0.78**	
		R-finetune	0.82	▽	**0.83**		0.81	▽	0.81	▽	0.77	▽	**0.78**	

Table 12: Highest accuracy for syntactical (obj), C-alignment (C-a), and N-alignment (N-a) heads in different models. $^{\times}$ means the accuracy is not high enough to be identified as function head. △ (▽) means finetuning increased (decreased) the head accuracy (i.e., R-finetune $>$ ($<$) R-multi).

accuracy of different types of function heads in the multilingual and bilingual models. As shown in Table 12, finetuning a multilingual model for a specific translation direction tended to increase accuracies of function heads (e.g., the syntactical obj heads and C-alignment heads) for the given translation direction, which is unsurprising. However, Table 12 indeed shows two interesting head behaviors in multilingual and bilingual models.

Cooperative Behavior of N-alignment Heads. First, we find the highest N-alignment accuracy tended to decrease instead of increasing during finetuning (the average accuracy of N-alignment heads also decreased). The fact that finetuning decreased N-alignment accuracies is surprising, considering N-alignment heads are crucial for predicting the next target word. We hypothesize that this is because N-alignment heads work in a cooperative way. Since there are multiple different gold translations (reorderings) in the target language for one source sentence, the Transformer uses different heads to learn different options for predicting (aligning) the next target word. Thus, the accuracy of a single N-alignment head is less important. Table 13 gives the accuracy of at least one head from the 4th layer (the most important N-alignment layer) correctly aligning the next target word. The results show that when we consider the whole layer, the next word alignment accuracy is fairly high and the layer accuracy generally increased during finetuning. The fact that the accuracy of any individual

	DeEn	FrEn	RoEn	EnDe	EnFr	EnRo
R-b	**96.8**	**97.2**	94.7	**96.7**	**95.9**	94.3
R-m	96.7	96.9	95.7	96.3	95.3	94.3
R-f	**96.8**	96.8	**96.0**	96.4	95.4	**95.2**

Table 13: Layer accuracy (4th layer) for N-alignment. R-b: R-bi; R-m: R-multi; R-f: R-finetune.

N-alignment head nevertheless tended to decrease during finetuning while the overall N-alignment layer accuracy tended to increase during finetuning indicates that N-alignment heads work in a cooperative way to collect different options for word reordering.

Multilingual Interference for Head Accuracy. Second, we find that, although finetuning generally improved the accuracies of function heads, the finetuned models (R-finetune) still tended to have lower accuracy than the bilingual baseline models (R-bi), especially the C-alignment accuracy and N-alignment layer accuracy for our high-resource language pairs De↔En and Fr↔En, as shown in Tables 12 and 13. This is surprising considering that R-finetune achieved higher translation qualities compared to R-bi, and suggests that language interference tends to cause an accuracy decrease for function heads and can be an important disadvantage of multilingual models compared to bilingual models. Regarding the reason why R-finetune generally had lower head accuracies but higher translation quality compared to R-bi for De↔En and Fr↔En tasks, it could be that the multilingual

pretraining helps the model to learn better representations (word embeddings) for less frequent words via the shared vocabulary.

4 Conclusion

This paper analyzes attention head functions and behaviors in multilingual Transformer translation models. We find that different language pairs, in spite of having different syntax and word orders, tend to share the same heads for the same functions, such as syntax heads and reordering heads. However, the different characteristics of different language pairs clearly cause interference in function heads and affect head accuracies, which can be an important disadvantage of multilingual models compared to bilingual models. Additionally, we reveal an interesting behavior of the Transformer cross-attention: the deep-layer cross-attention heads work in a cooperative way to learn different options for word reordering, which can be caused by the nature of translation tasks having multiple different gold translations (reorderings) in the target language for one source sentence.

Limitations

Our study focuses on models trained for particular source to target language pairs. It covers six translation directions with limited typological diversity in the considered languages, due to the need for ground truth word alignments. In future work, multilingual models covering many more languages with more linguistic diversity can be investigated following our methodology.

Acknowledgements

The authors acknowledge the financial support by the German Federal Ministry for Education and Research (BMBF) within the project "KI-Servicezentrum Berlin Brandenburg" 01IS22092. Hongfei Xu is supported by the National Natural Science Foundation of China (Grant No. 62306284) and the Natural Science Foundation of Henan Province (Grant No. 232300421386).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Mike Schuster, Noam Shazeer, Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2018. The best of both worlds: Combining recent advances in neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–86, Melbourne, Australia. Association for Computational Linguistics.

Yun Chen, Yang Liu, Guanhua Chen, Xin Jiang, and Qun Liu. 2020. Accurate word alignment induction from neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 566–576, Online. Association for Computational Linguistics.

Ting-Rui Chiang, Yi-Pei Chen, Yi-Ting Yeh, and Graham Neubig. 2022. Breaking down multilingual machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2766–2780, Dublin, Ireland. Association for Computational Linguistics.

Javier Ferrando and Marta R. Costa-jussà. 2021. Attention weights in transformer NMT fail aligning words between sequences but largely explain model predictions. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 434–443, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Zae Myung Kim, Laurent Besacier, Vassilina Nikoulina, and Didier Schwab. 2021. Do multilingual neural machine translation models contain language pair specific attention heads? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*,

pages 2832–2841, Online. Association for Computational Linguistics.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. Investigating multilingual NMT representations at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1565–1575, Hong Kong, China. Association for Computational Linguistics.

Zehui Lin, Liwei Wu, Mingxuan Wang, and Lei Li. 2021. Learning language specific sub-network for multilingual machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 293–305, Online. Association for Computational Linguistics.

Christos Louizos, Max Welling, and Diederik P Kingma. 2017. Learning sparse neural networks through l_0 regularization. *arXiv preprint arXiv:1712.01312*.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.

Telmo Pires, Robin Schmidt, Yi-Hsiu Liao, and Stephan Peitz. 2023. Learning language-specific layers for multilingual machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14767–14783, Toronto, Canada. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Uri Shaham, Maha Elbayad, Vedanuj Goswami, Omer Levy, and Shruti Bhosale. 2023. Causes and cures for interference in multilingual translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15849–15863, Toronto, Canada. Association for Computational Linguistics.

Gongbo Tang, Philipp Rönchen, Rico Sennrich, and Joakim Nivre. 2021. Revisiting negation in neural machine translation. *Transactions of the Association for Computational Linguistics*, 9:740–755.

Gongbo Tang, Rico Sennrich, and Joakim Nivre. 2018. An analysis of attention mechanisms: The case of word sense disambiguation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 26–35, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019a. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4396–4406, Hong Kong, China. Association for Computational Linguistics.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019b. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Yining Wang, Long Zhou, Jiajun Zhang, Feifei Zhai, Jingfang Xu, and Chengqing Zong. 2019. A compact and language-sensitive multilingual translation method. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1213–1223, Florence, Italy. Association for Computational Linguistics.

Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, Online. Association for Computational Linguistics.

Wanying Xie, Yang Feng, Shuhao Gu, and Dong Yu. 2021. Importance-based neuron allocation for multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for*

Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 5725–5737, Online. Association for Computational Linguistics.

Haoran Xu, Jean Maillard, and Vedanuj Goswami. 2023. Language-aware multilingual machine translation with self-supervised learning. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 526–539, Dubrovnik, Croatia. Association for Computational Linguistics.

Hongfei Xu, Qiuhui Liu, Josef van Genabith, and Deyi Xiong. 2021a. Modeling task-aware MIMO cardinality for efficient multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 361–367, Online. Association for Computational Linguistics.

Hongfei Xu, Josef van Genabith, Qiuhui Liu, and Deyi Xiong. 2021b. Probing word translations in the transformer and trading decoder for encoder layers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 74–85, Online. Association for Computational Linguistics.

Fei Yuan, Yinquan Lu, Wenhao Zhu, Lingpeng Kong, Lei Li, Yu Qiao, and Jingjing Xu. 2023. Lego-MT: Learning detachable models for massively multilingual machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11518–11533, Toronto, Canada. Association for Computational Linguistics.

Biao Zhang, Ankur Bapna, Rico Sennrich, and Orhan Firat. 2021. Share or not? learning to schedule language-specific capacity for multilingual translation.

Jingyi Zhang and Josef van Genabith. 2021. A bidirectional transformer based alignment model for unsupervised word alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 283–292, Online. Association for Computational Linguistics.

Changfeng Zhu, Heng Yu, Shanbo Cheng, and Weihua Luo. 2020. Language-aware interlingua for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1650–1655, Online. Association for Computational Linguistics.

Yaoming Zhu, Jiangtao Feng, Chengqi Zhao, Mingxuan Wang, and Lei Li. 2021. Counter-interference adapter for multilingual machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2812–2823, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bridging the Gap Between Position-Based and Content-Based Self-Attention for Neural Machine Translation

Felix Schmidt
AppTek
Aachen, Germany
fschmidt@apptek.com

Mattia Antonino Di Gangi
AppTek
Aachen, Germany
mdigangi@apptek.com

Abstract

Position-based token-mixing approaches, such as FNet and MLPMixer, have shown to be exciting attention alternatives for computer vision and natural language understanding. The motivation is usually to remove redundant operations for higher efficiency on consumer GPUs while maintaining Transformer quality. On the hardware side, research on memristive crossbar arrays shows the possibility of efficiency gains up to two orders of magnitude by performing in-memory computation with weights stored on device. While it is impossible to store dynamic attention weights based on token-token interactions on device, position-based weights represent a concrete alternative if they only lead to minimal degradation. In this paper, we propose position-based attention as a variant of multi-head attention where the attention weights are computed from position representations. A naive replacement of token vectors with position vectors in self-attention results in a significant loss in translation quality, which can be recovered by using relative position representations and a gating mechanism. We show analytically that this gating mechanism introduces some form of word dependency and validate its effectiveness experimentally under various conditions. The resulting network, rPosNet, outperforms previous position-based approaches and matches the quality of the Transformer with relative position embedding while requiring 20% less attention parameters after training.[1]

1 Introduction

The Transformer (Vaswani et al., 2017) revolutionized the field of neural machine translation before its wide adoption in numerous other tasks (Dong et al., 2018; Devlin et al., 2019; Chen et al., 2021; Dosovitskiy et al., 2021). Using self-attention (Vaswani et al., 2017), the Transformer computes high-level representations for each token as a weighted sum of the entire sequence, where the weights depend on the pairwise content interactions. However, recent work argues that results similar to the Transformer can also be achieved by modeling self-attention weights based on positional instead of content information (Wu et al., 2019; You et al., 2020; Tolstikhin et al., 2021; Liu et al., 2021; Lee-Thorp et al., 2022). Often, these position-based methods are used with some form of gating mechanism that precedes or wraps the token-mixing operation (Wu et al., 2019; Liu et al., 2021; Kim et al., 2023).

Position-based self-attention alternatives often speed up the computation on commercial computing devices like GPU, but they can become more attractive from the perspective of using memristive crossbar arrays (Chua, 1971; Strukov et al., 2008). Recent advances in analog in-memory computation with memristive crossbar arrays have shown impressive efficiency improvements in the inference of deep learning models (Hu et al., 2018; Wang et al., 2019; Kataeva et al., 2019; Yao et al., 2020; Xue et al., 2021), up to 110 times better energy efficiency and 30 times better performance density compared to a Tesla V100 GPU (Yao et al., 2020). However, such efficiency is obtained by storing weights of matrix-vector multiplications in the device rather than calculating them on the fly, which excludes the possibility of using attention to compute the weight matrix.

With the goal of finding self-attention alternatives for machine translation that can be more easily used with memristive crossbar arrays, we compare existing position-based approaches and observe a significant quality loss when they use no form of gating. Additionally, by scoring with a diverse set of metrics, we show that, even with gating, no existing approach can consistently match Transformer results. While the role of gating to guide the information flow of neural networks is known (Srivastava et al., 2015; Dauphin et al., 2017), its

[1] Code available at https://github.com/apptek/posnet-position_based_attention

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 507–521
December 6–7, 2023.

importance for the performance of position-based approaches has yet to be explored.

In this paper, we propose aPosNet and rPosNet, two position-based networks that leverage gating and compute self-attention weights based on the interactions of **a**bsolute and **r**elative position representations. Both differ slightly from the Transformer baseline with relative position embeddings (Shaw et al., 2018), which enables us to deliver insights into gating and its dependency on position information. In summary, we provide the following contributions:

- Analytically, we derive that wrapping the weighted sum of tokens with a gating mechanism introduces latent content-dependent token-mixing weights (Section 3).
- We provide an inference-time matrix precomputation for positional attention that can be easily stored in device (Section 4).
- rPosNet outperforms existing position-based methods and performs on par with the Transformer with relative position embeddings while saving 20% of the self-attention parameters (Section 6).
- We show that increasing the expressiveness of token-mixing weights reduces the usefulness of gating, coherently with the idea that it enables content-based interactions (Section 7.1).
- We observe experimentally that rPosNet is less effective when used in cross-attention. Our gating reformulation suggests one probable reason, but we leave detailed investigations for future work (Section 7.2).

2 Background

Neural machine translation is typically modeled with an encoder-decoder sequence-to-sequence (Sutskever et al., 2014) Transformer, which mainly consists of multi-head attention and feed-forward sub-layers. In the following, we introduce our notation, position-based token-mixing alternatives and the gating mechanism commonly used in modern architectures.

2.1 Multi-head attention

Given a source sequence representation $\mathbf{x} \in \mathbb{R}^{M \times D}$ and target sequence representation $\mathbf{y} \in \mathbb{R}^{N \times D}$, the multi-head attention mechanism (Vaswani et al., 2017) mixes the elements in $\mathbf{x}$ for every element in $\mathbf{y}$. If $\mathbf{y}$ and $\mathbf{x}$ refer to the same sequence, it is called self-attention. The multi-head concept derives from performing the following operations on H parallel splits of the feature dimension D. In this work, we drop the head indices for simplicity of notation. To calculate the unnormalized mixing weight, referred to as attention energy, of y_n and x_m, those are projected into query and key and combined using the dot product:

$$\hat{\alpha}_{nm} := \frac{(W^Q y_n)(W^K x_m)^\top}{\sqrt{D}}. \tag{1}$$

Since $\hat{\alpha}_{nm}$ is computed from token contents, we say that attention captures token-token interactions. The attention weight is then calculated by the softmax normalization of the attention energy:

$$\alpha_{nm} := \frac{\exp \hat{\alpha}_{nm}}{\sum_{m'} \exp \hat{\alpha}_{nm'}}, \tag{2}$$

and used as the token-mixing weight in the weighted sum over projected input tokens $\mathbf{x}$, denoted value vectors:

$$c_n := \sum_m \alpha_{nm} \cdot (W^V x_m). \tag{3}$$

We will refer to the result c_n as context vector. Finally, the context vectors of each head are concatenated and mixed with a linear projection, called output projection.

2.2 Position-based token mixing

We briefly overview how existing position-based token-mixing approaches propose to modify the attention weights and provide the corresponding Equations in Appendix A for comparison.

FNet Proposed for language understanding, FNet (Lee-Thorp et al., 2022) applies a 2D Fourier transform over the spatial and feature dimension of $\mathbf{x}$. However, this formulation performed poorly in preliminary experiments, which is why our FNet implementation, denoted FourierNet, applies a 1D Fourier transform along the spatial dimension and employs value and output projections. We will show in our results that, despite its claimed good quality for natural language understanding, the translation quality achieved by FourierNet is significantly lower than Transformer.

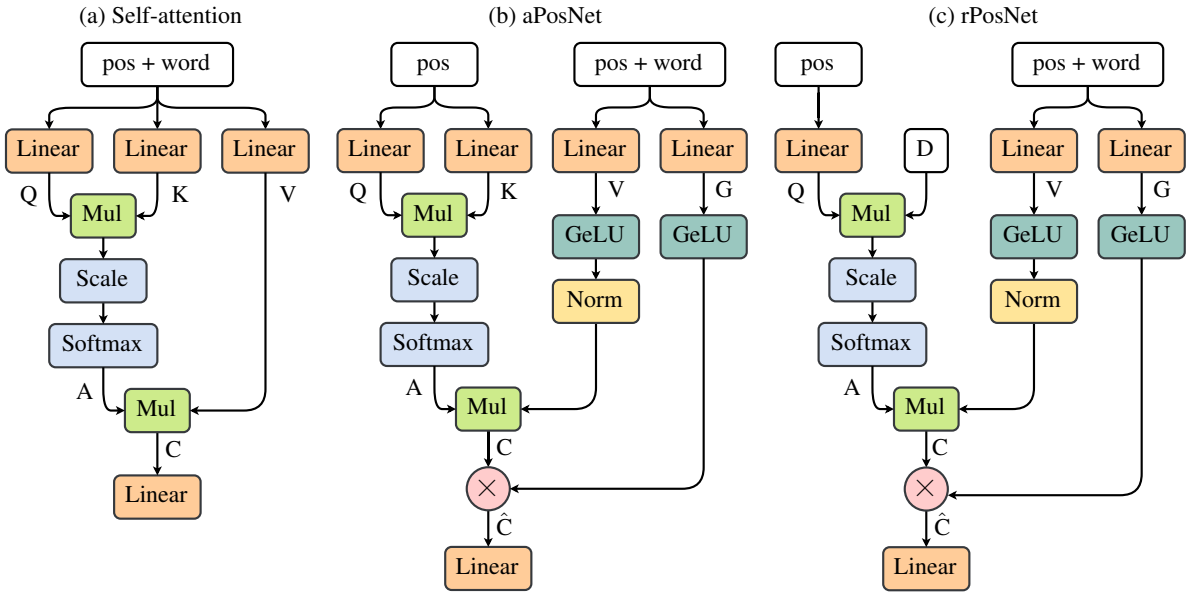


Figure 1: Flowchart representation of (a) self-attention, (b) gated absolute position-based attention (aPosNet), and (c) gated relative position-based attention (rPosNet). While self-attention provides word and position information to queries (Q) and keys (K), we omit word information to calculate the attention weights of PosNet. In rPosNet, we model relative positions using relative position representations (D). In addition, we employ the gating mechanism presented in Section 2.3, which applies GeLU activation and layer normalization on the values (V) and elementwise multiplies the context vector (C) with the GeLU activated gate (G) resulting in the gated context vector (Ĉ).

GaussianNet Proposed for machine translation, You et al. (2020) hardcode self-attention weights as a Gaussian distribution. They report similar performance to the Transformer when GaussianNet is applied for self-attention but a significant degradation if extended to cross-attention.

LinearNet Tolstikhin et al. (2021) propose mixing tokens with a learnable spatial projection, effectively representing α. It has been proposed, together with other architecture changes, for image classification and natural language understanding with minor degradations to the Transformer.

LightConv For machine translation and other tasks, Wu et al. (2019) introduce a lightweight form of depthwise convolution, which shares the kernel weights W across the feature dimension of a head and the outputs while additionally softmax normalizing them.

gLinearNet Liu et al. (2021) combine the spatial projection of LinearNet with the gating mechanism of Section 2.3. They propose their architecture for image classification and masked language modeling and report significant improvements over Tolstikhin et al. (2021).

2.3 Gating mechanisms

Various formulations of gating mechanisms have been proposed to control the information flow in neural networks (Hochreiter and Schmidhuber, 1997; Cho et al., 2014; Srivastava et al., 2015; van den Oord et al., 2016; Dauphin et al., 2017). They all have in common an elementwise multiplication between two vectors where one, the gate, is bounded in the $[0, 1]$ interval. The gating mechanism we consider here has been proven effective with position-based token-mixing approaches (Liu et al., 2021; Kim et al., 2023) and differs from other gating mechanisms in that the gate is GeLU activated (Hendrycks and Gimpel, 2016) and thus only lower bounded. This gating mechanism modifies the weighted sum of Equation 3 by applying layer normalization (Ba et al., 2016) on the value vector $v_m = W^V x_m$ and elementwise multiplying the context vector with the gate $g_n = \sigma_g(W^G y_n)$:

$$\hat{c}_n := \Big[\sum_m \alpha_{nm} \cdot \mathrm{Norm}\big(\sigma_g(v_m)\big)\Big] \odot g_n, \quad (4)$$

where σ_g refers to the GeLU function and $\hat{c}_n$ to the gated context vector. In general, gating can be applied with any formulation of α. However, we will show experimentally in Section 7.1 that its benefits

strongly depend on the information incorporated within α.

3 Reformulating the Gating Mechanism

To better understand the implications of gating, we reformulate Equation 4. We omit layer normalization for simplicity and will show in the Appendix B that the general reformulation is unaffected if we apply layer normalization on v_m. Additionally, we leverage the GeLU approximation $\sigma_g(v_m) \approx v_m \sigma_s(1.702 v_m)$, where σ_s refers to the Sigmoid function, and rewrite Equation 4 as

$$\hat{c}_n \approx \sum_m \alpha_{nm} \beta_{nm} \odot v_m. \tag{5}$$

Equation 5 shows that gating the context vector introduces the latent weights $\beta_{nm} \in \mathbb{R}^D$:

$$\beta_{nm} = g_n \odot \sigma_s(1.702 v_m), \tag{6}$$

which consists of the two independent factors $\beta'_n = g_n$ and $\beta'_m = \sigma_s(1.702 v_m)$. While the multiplication of β'_n and β'_m, in general, allows for token-token interactions, the independence of these factors poses a limitation: for a given query token y_n, the ratio between the weights assigned to x_m and to $x_{m'}$ is independent of y_n:

$$\frac{\beta_{nm}}{\beta_{nm'}} = \frac{\beta'_m}{\beta'_{m'}}. \tag{7}$$

In other words, the ratios of token-mixing weights for a query y_n as computed by β are predetermined by the ratios across $\beta'_{1\ldots M}$. While we show in Section 6 that this limitation is not problematic for self-attention, it may be part of the reason gating and relative position-based attention are not effective in cross-attention (see Section 7.2).

4 Position-based Attention

In this Section, we propose position-based attention, which determines the token-mixing weight connecting tokens x_m and y_n solely based on the position-position interactions between n and m. We pair position-based attention with the gating mechanism of Section 2.3.

4.1 Absolute position-based attention

In absolute position-based attention we compute the attention energy as the dot product between the two projected position embeddings $\tilde{n}$ and $\tilde{m}$:

$$\hat{\alpha}_{nm} := \frac{(W^Q \tilde{n})(W^K \tilde{m})^\top}{\sqrt{D_h}}. \tag{8}$$

While $\tilde{n}$ and $\tilde{m}$ are shared across all layers, W^Q and W^K are layer-specific. We refer to the combination of Equation 8 and the gating mechanism of Section 2.3 with aPosNet.

Pre-computing the attention energies The query and key inputs $\tilde{n}$ and $\tilde{m}$ are independent of the word representations y_n and x_m and are constant after training. Since the attention energy values $\hat{\alpha}_{nm}$ only depend on $\tilde{n}$ and $\tilde{m}$, we can pre-compute $\hat{\alpha}$ and obtain a matrix of the form $(H \times N \times M)$ that can be used during inference. In the following theoretical complexity discussions we set $N = M$ for simplicity of notation.

Theoretical complexity Apart from the gating overhead, aPosNet has similar theoretical complexity as attention. However, by pre-computing $\hat{\alpha}$, we can skip the dot-product and key query projections, reducing[2] the number of parameters from $5D^2$ to $HN^2 + 3D^2$ and the number of operations from $2N^2D + 5ND^2$ to $N^2D + 3ND^2$. We compare theoretical complexities in Appendix C.

Relation to gLinearNet With the pre-computed attention energy matrix, aPosNet becomes similar to gLinearNet except that α of gLinearNet is not normalized and has been trained directly.

4.2 Relative position-based attention

To model position interactions with relative position-based attention, we borrow the relative position representations $\tilde{d}_{nm}$ from Shaw et al. (2018), which we use in the dot product with the projected position embedding $\tilde{n}$:

$$\hat{\alpha}_{nm} := \frac{(W^Q \tilde{n})(\tilde{d}_{nm})^\top}{\sqrt{D_h}}. \tag{9}$$

Similarly to Shaw et al. (2018), the distance embedding $\tilde{d}$ is clipped to a maximum unidirectional context size K:

$$\tilde{d}_{nm} := \text{Embedding}_{\text{rel}}\Big(\text{clip}\big(\lfloor \gamma n \rfloor - m, K\big)\Big). \tag{10}$$

However, in contrast to Shaw et al. (2018), we extend relative position-based self-attention to be compatible with cross-attention by multiplying n with the length ratio $\gamma := \frac{M}{N}$ which we determine similar to You et al. (2020) by measuring the average length ratio on the training set. We refer to the

[2] Typically in sentence-level machine translation we have $N \ll D$.

Table 1: Dataset statistics.

Dataset	Vocab. Size		Train Pairs	Test Pairs	Valid Pairs
	Src	Tgt			
DE→EN	10k		160k	6750	7283
EN→DE	44k		4M	3003	40k
EN→FR	46k		36M	3003	27k
EN→ZH	32k	45k	17M	2001	13k

combination of Equation 9 and the gating mechanism of Section 2.3 with rPosNet. In Figure 1, we illustrate the operations performed by aPosNet and rPosNet in comparison to multi-head self-attention.

Pre-computing the attention energies Similar to aPosNet, we can pre-compute $\hat{\alpha}$ of rPosNet after training, which summarizes the interactions between query and relative position representations into a matrix of shape $(H \times \hat{K} \times N)$, where $\hat{K} = 2K + 1$. While the attention energy matrix of aPosNet grows quadratically with the length of the sequence, rPosNet's matrix grows linearly due to the constant size $\hat{K}$ of the relative position representations.

Theoretical complexity Pre-computing $\hat{\alpha}$ after training reduces the number of parameters from $\hat{K}D + 4D^2$ to $H\hat{K}N + 3D^2$ and operations from $\hat{K}ND + N^2D + 4ND^2$ to $N^2D + 3ND^2$. Inserting the Base model configuration of Section 5 ($D = 2048$, $\hat{K} = 33$) and the maximum sentence length $N = 128$, this pre-computation of $\hat{\alpha}$ saves 23% of attention parameters.

Relation to LightConv After pre-computing $\hat{\alpha}$, rPosNet differs from LightConv in that rPosNet's weights have global context and depend also on the absolute query position, and as such are not shared across y_n. We provide an ablation study in Section 7.3 to understand the importance of these differences.

5 Experimental Setup

5.1 Datasets & evaluation

We perform our comparison on four datasets of varying sizes: IWSLT14 German-English (Federico et al., 2014), WMT14 English-{German, French} (Bojar et al., 2014), and WMT18 English-Chinese (Bojar et al., 2018). We split each dataset into train and validation pairs and evaluate DE→EN models on the test sets TED-{dev10,dev12, test10, tst11, tst12}, EN-{DE, FR} models on newstest14 and EN→ZH models on newstest17. An overview of the dataset statistics is shown in Table 1. We preprocess all datasets using Byte Pair Encoding (BPE) (Sennrich et al., 2016) and lowercase the text for the DE→EN direction.

We report BLEU (Papineni et al., 2002), BLEURT (Sellam et al., 2020), and COMET (Rei et al., 2020) for each evaluation. All scores are calculated on detokenized text. To calculate BLEU scores, we use sacreBLEU[3] and its internal tokenizations[4][5]. For BLEURT and COMET, we use the official implementations[6][7] and the models *BLEURT-20* and *wmt20-comet-da*, respectively. To summarize results, we will refer to the translation quality difference between two approaches as their relative difference averaged across all metrics and datasets.

5.2 Model architectures

Our Base and Big Transformer architectures follow the implementation of Vaswani et al. (2017), whereas, for the Small models, we halve the feed-forward dimension to 1024 and increase dropout to 0.3. We compare position-based token-mixing approaches by leveraging the respective formulations instead of encoder/decoder self-attention while leaving the rest of the Transformer architecture unchanged. We make an exception for FourierNet, which cannot be straightforwardly extended to the decoder because it has an explicit dependency on the sequence length. Instead, FourierNet uses multi-head attention within decoder self-attention.

In preliminary experiments, we found that aPosNet works best with sinusoidal positional embeddings (Vaswani et al., 2017) and rPosNet with learnable embeddings (Gehring et al., 2017). All other position-based token-mixing approaches use sinusoidal positional embeddings. Similar to Shaw et al. (2018), our implementation of rPosNet and LightConv use a unidirectional context window $K = 16$ for the Base and $K = 8$ for the Big model.

5.3 Training setup

Our training setup closely follows the configuration of Vaswani et al. (2017). Similarly, we use

[3] https://github.com/mjpost/sacrebleu

[4] SacreBLEU signature for EN, FR, DE: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.0.0

[5] SacreBLEU signature for ZH: nrefs:1|case:mixed|eff:no|tok:zh|smooth:exp|version:2.0.0

[6] https://github.com/google-research/bleurt

[7] https://github.com/Unbabel/COMET

Table 2: Base model results on EN→DE, EN→FR, and EN→ZH. We calculate statistical significance (p-value $\leqslant 0.05$) using paired bootstrap resampling with respect to the Transformer (†) and to Shaw et al. (2018) (‡). Note that this scoring differs from Vaswani et al. (2017) in that they split German compound words, which usually increases the BLEU score, and from You et al. (2020) in that we use sacreBLEU's default tokenizer, not 'intl'. We ensured that our baseline system and reimplementation of You et al. (2020) match in BLEU when evaluating similarly.

Model	Params	EN→DE			EN→FR			EN→ZH		
	(EN→DE)	BLEU	BLEURT	COMET	BLEU	BLEURT	COMET	BLEU	BLEURT	COMET
Transformer	66.5M	26.3	71.1	47.6	37.8	69.0	61.1	33.8	64.3	42.5
Shaw et al. (2018)	66.7M	26.3	**71.4**	**48.6**	37.8	69.2	61.6	**34.0**	**64.6**	**43.5**
FourierNet	63.4M	22.8†‡	66.0†‡	31.8†‡	34.9†‡	64.2†‡	49.3†‡	31.5†‡	61.6†‡	34.9†‡
GaussianNet	60.2M	25.3†‡	68.1†‡	39.5†‡	36.7†‡	66.9†‡	55.7†‡	32.6†‡	62.6†‡	36.8†‡
LinearNet	61.8M	25.3†‡	69.8†‡	44.3†‡	37.0†‡	67.7†‡	58.2†‡	33.1†‡	63.3†‡	40.2†‡
LightConv	63.4M	26.0†‡	70.6†‡	46.7†‡	37.4†‡	68.6†‡	60.3†‡	33.0†‡	63.5†‡	41.1†‡
gLinearNet	65.0M	26.1	70.8‡	46.7‡	37.8	69.1	61.3	33.5‡	64.0‡	42.4‡
aPosNet	65.0M	25.9†‡	70.6†‡	46.1†‡	37.7	69.0	61.4	33.6‡	63.7‡	42.2‡
rPosNet	63.9M	**26.6**	**71.4**†	**48.6**	**37.9**	**69.4**†	**61.8**	33.8	64.2	43.1

Table 3: Big model results on EN→DE and Small model results on DE→EN.

Model	EN→DE				DE→EN			
	Params	BLEU	BLEURT	COMET	Params	BLEU	BLEURT	COMET
Transformer	221M	27.1	72.3	50.4	36.8M	35.0	69.3	37.6
Shaw et al. (2018)	221M	**27.3**	**72.7**†	**51.5**†	37.0M	**35.4**†	**69.7**†	**38.8**†
FourierNet	208M	24.0†‡	67.6†‡	36.5†‡	33.6M	32.5†‡	66.9†‡	28.2†‡
GaussianNet	196M	26.3†‡	69.4†‡	42.3†‡	30.4M	34.3†‡	68.4†‡	34.1†‡
LinearNet	199M	26.6†‡	71.3†‡	48.0†‡	32.0M	34.0†‡	68.3†‡	33.9†‡
LightConv	209M	26.8†‡	71.7†‡	49.1†‡	33.6M	34.4†‡	68.9†‡	35.5†‡
gLinearNet	212M	27.1	72.2‡	49.9‡	35.2M	34.5†‡	69.0†‡	36.3†‡
aPosNet	212M	26.8†‡	71.4†‡	47.7†‡	35.2M	34.2†‡	68.5†‡	34.7†‡
rPosNet	210M	**27.3**	72.2‡	50.4‡	34.1M	35.1‡	69.5‡	38.2‡

the Adam optimizer (Kingma and Ba, 2014) and a warmup learning rate schedule with 4000 steps. We group batches by sentence length and train the Small models for 30k steps, the Base models for 150k, and the Big models for 300k.

The final model is an average over the best checkpoint and its following if there are enough checkpoints to average, or else we take an average over the last checkpoints. We determine the best checkpoint by its perplexity on the validation set. For DE→EN, we consistently average 30 checkpoints with a checkpoint period of 300 steps; for the Base models, we average 7 checkpoints with 1000 steps each; for the Big models, 20 checkpoints with 600 steps each.

The Small models use an effective batch size of approximately 16000 target tokens while the Base and Big models accumulate approximately 27000 target tokens per step. The source and target sentence lengths are restricted to 128 tokens. We use beam search with a beam size of 12 for all models. All models in this work are implemented in PyTorch (Paszke et al., 2019). The Small models are trained on a single 2080 TI graphics card, the Base models on two, and the Big models on four.

6 Results

We compare translation quality of the Base model configurations in Table 2, and Small and Big model configurations in Table 3.

Gated position-based attention In all experiments, we observe rPosNet performing as well or slightly better than the Transformer with an average translation quality increase of 0.7% across

all test sets and metrics. It shows that the self-attention weights of rPosNet, consisting of content-dependent β and position-dependent α, achieve sufficient expressiveness for machine translation. aPosNet cannot match this expressiveness and underperforms the Transformer with an average relative degradation of 1.8%. In the Small setup on DE→EN, this reaches an absolute degradation of 2.9 points in COMET and 0.8 points in BLEURT. The significant difference between aPosNet and rPosNet highlights the importance of relative position information in α.

The results of gLinearNet and LightConv further emphasize the strong modeling capabilities of absolute (query) and relative position (key) interactions in rPosNet. In comparison, token-mixing weights in gLinearNet solely model absolute position interactions and in LightConv relative position interactions. Both cannot match rPosNet's translation quality, with gLinearNet on average lacking behind by 1.3% relative and LightConv by 2.4%. Note that in contrast to Wu et al. (2019), we do not match parameters between LightConv and the Transformer. Most prominent in the Base setting on EN→DE, rPosNet outperforms gLinearNet by 0.6 BLEURT and 1.9 COMET points. While aPosNet cannot match Transformer results, rPosNet consistently outperforms other position-based methods and is on par with Shaw et al. (2018) and the Transformer across most model sizes and data conditions.

Hard-coded token-mixing weights Our results show that hard coding encoder self-attention weights as the twiddle factors of the Fourier transform (FourierNet) leads to poor results for machine translation and, on average across all datasets and metrics, degrades translation quality relative to the Transformer by 13.2%. In GaussianNet, weights are manually designed to follow the normal distribution of Transformer self-attention patterns, which significantly reduces the degradation to 6.3%. However, the translation quality is still considerably worse than LinearNet's, the weakest model with trainable self-attention weights. The difference between LinearNet and GaussianNet is negligible in BLEU but made visible with BLEURT and COMET, which correlate better with human judgment (Kocmi et al., 2021). In particular, we confirmed by manually analyzing a sample of translations (see Appendix E) that the semantic metrics discriminate better between translation hypotheses when they all have little overlap with the references

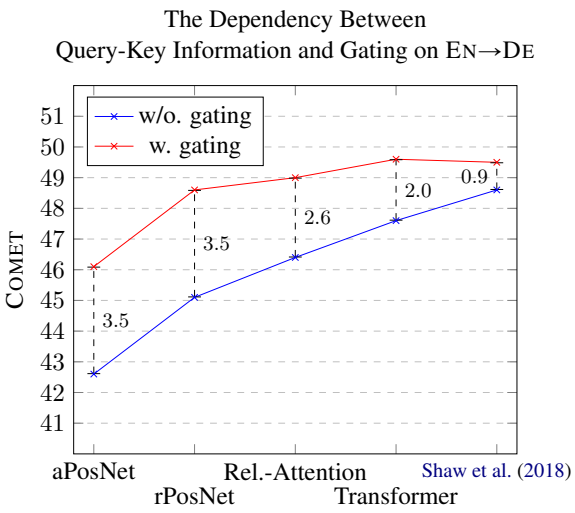


Figure 2: Approaches depicted on the x-axis differ in the provided information to queries and keys. On the y-axis we depict COMET, which is the most accurate metric according to Kocmi et al. (2021), and provide the full Table showing BLEU, BLEURT, and COMET in Appendix D. If position information is provided to queries and keys, gating has a significant positive impact on translation quality that diminishes with the usage of content information.

or changing a single word alters the meaning of the sentence. Thus, approaches with learnable token-mixing weights, such as rPosNet, are considerably better than hard-coded approaches.

7 Analysis

7.1 The impact of gating and query-key information

The gating mechanism is known to guide the learning of cross-token patterns (Tu et al., 2017; Dauphin et al., 2017). In Section 3, we mathematically showed that by gating the context vector, these patterns are captured within the latent token-mixing weights β. Since the products of β and α form the actual token-mixing weights, we analyze in this Section how content information in α impacts the usefulness of gating. For that, we compare the utilization of position versus content information in the query and key input of self-attention, with and without gating. The results are visualized in Figure 2, where we depict COMET scores on the y-axis and the query and key input on the x-axis.

The formulation of position-based attention without gating primarily[8] differs from the Transformer in the provided information within queries and

[8]The position embeddings may also differ between approaches.

keys. Relative attention uses the relative position representations of Shaw et al. (2018)'s approach but without the query-key dot product of multi-head attention. Thus, relative attention differs from rPosNet in that content information is provided to the queries and is equal to dynamic convolutions (Wu et al., 2019) with global context (Chang et al., 2021). In total, the x-axis of Figure 2 depicts position-position interactions for aPosNet and rPosNet, token-position interactions for relative attention, token-token interactions for the Transformer, and token-token + token-position interactions for Shaw et al. (2018). We sort these approaches on the x-axis in order of their attention weight expressiveness.

Figure 2 shows that the gating mechanism of the position-based attention approaches aPosNet and rPosNet increases COMET by 3.5 points. On the other hand, content-based approaches leverage gating with a lower absolute COMET increase of 2.6 points for relative attention, 2 points for the Transformer and only 0.9 points for Shaw et al. (2018). Thus, gating is less helpful if α can capture content-dependent patterns, and increasing the expressiveness of those patterns diminishes the usefulness of gating. Since gating introduces an additional projection matrix of size D^2 per self-attention layer, content-based mixing approaches may just leverage the additional parameters, but we leave further investigation for future work. In contrast, approaches that do not incorporate content information within the attention weights can benefit from token-token interactions captured in β. Additionally, the comparable performance of rPosNet and relative attention with gating suggests that gating makes the content information within relative attention redundant for translation quality.

7.2 Comparing the usage of rPosNet across attention layers

While the aforementioned experiments concentrated on self-attention, we also consider cross-attention in this Section and analyze how the usage of rPosNet affects translation quality compared to multi-head attention. In Table 4, we depict the translation quality on EN→DE when combinations of encoder self-attention (enc-self), decoder self-attention (dec-self), and decoder cross-attention (dec-cross) employ multi-head attention (✗) or rPosNet (✓). The model using rPosNet only for cross-attention while all other layers employ multi-head attention (row 4) significantly decreases translation quality by 5.7% relative to the Transformer. The result suggests that content-dependent patterns incorporated by β cannot sufficiently capture source-target token interactions. We hypothesize that part of the reason is the inability of β to express varying relations across source tokens (see Section 3). While this may be a significant limitation of gating, we leave the exploration of this and other possible reasons to future work.

Table 4: A translation quality comparison of all combinations in which encoder self-attention (enc-self), decoder self-attention (dec-self), and/or decoder cross-attention (dec-cross) use either multi-head attention (✗) or rPosNet (✓). We conduct the experiments on EN→DE and report BLEU, BLEURT, and COMET.

rPosNet Layers			EN→DE		
enc-self	dec-self	dec-cross	BLEU	BLEURT	COMET
✗	✗	✗	26.3	71.1	47.6
✓	✗	✗	26.4	71.2	48.1
✗	✓	✗	26.1	71.1	47.2
✗	✗	✓	24.6	69.2	43.8
✓	✓	✗	**26.6**	**71.4**	**48.6**
✓	✗	✓	24.8	69.8	45.2
✗	✓	✓	24.3	69.0	42.8
✓	✓	✓	24.9	69.3	43.5

However, utilizing rPosNet within all self-attention layers (row 8), so that rPosNet is the only token-mixing method, does not lead to further degradation of translation quality with a relative degradation to the Transformer of 5.5% (5.3% relative in BLEU). Although the loss is substantial, rPosNet improves upon You et al. (2020)'s relative BLEU degradation of 12.3%[9]. Additionally, Table 4 shows that using rPosNet within decoder self-attention is only beneficial if encoder self-attention leverages rPosNet, whereas the usage within encoder self-attention always positively impacts translation quality.

7.3 From LightConv to rPosNet

With the similarities between LightConv and rPosNet, we want to understand what features of rPosNet are responsible for its better translation quality. While Wu et al. (2019) propose LightConv initially

[9]As reported by You et al. (2020)

Table 5: Starting from LightConv and progressively implementing the features of rPosNet.

Model	Params	EN→ZH		
		BLEU	BLEURT	COMET
Light Convolution	101M	33.2	63.4	40.6
+ GLU [LightConv]	104M	33.0	63.5	41.1
+ GeLU Gating	104M	33.5	63.7	42.4
+ Global Context	104M	33.6	63.9	42.4
rPosNet	104M	**33.8**	**64.2**	**43.1**

with the GLU mechanism (Dauphin et al., 2017) (see Equation 14), we differentiate between LightConv with and without GLU since the effect of gating is a central component of our analysis. We start with LightConv without GLU, denoted Light Convolution, and progressively implement the features of rPosNet. In Table 5, we show the translation quality on EN→ZH of the models leveraging the respective position-based approach instead of self-attention. Light Convolution (row 1) shows similar translation quality to LightConv (row 2). Replacing GLU gating with the gating mechanism of Section 2.3, denoted GeLU gating (row 3), increases translation quality noticeably by 0.5 points in BLEU, 0.2 points in BLEURT, and 1.3 points in COMET. Additionally, adding global context (row 4) by spreading the outer kernel weights across the whole sequence increases translation quality slightly by 0.1 BLEU and 0.2 BLEURT (no improvement in COMET). The remaining difference to rPosNet (row 5) is the different training scheme and rPosNet's unshared kernel weights across query positions. Together they add additional 0.2 points in BLEU, 0.3 in BLEURT, and 0.7 in COMET. The results show that all differences between LightConv and rPosNet are responsible for their translation quality difference. While the global context seems negligible for machine translation, GeLU gating, training scheme, and unshared token-mixing weights are the most important.

8 Related Work

The question of how to represent position and integrate it into the Transformer architecture has been a vast research field that we briefly want to overview and connect to our approach. An extensive line of research focuses on improving position embeddings (Kitaev et al., 2020; Liu et al., 2020; Kiyono et al., 2021) and their integration into the word vectors (Neishi and Yoshinaga, 2019; Wang et al., 2020). This direction is mainly orthogonal to our approach, and many ideas and methods can be leveraged with position-based attention. We leave these investigations for future work and restricted to learnable (Gehring et al., 2017) and sinusoidal (Vaswani et al., 2017) embeddings.

A different line of research focuses on integrating position within the attention mechanism (Shaw et al., 2018; Dai et al., 2019; Dufter et al., 2020; Huang et al., 2020; Raffel et al., 2020; Ke et al., 2020; He et al., 2021; Wu et al., 2021). They all improve over Transformer models for various tasks by modifying word and position interactions within the attention matrix and introducing relative position representations as a scalar or vector. While they still rely on content-dependent attention weights, they showed the importance of relative position representations, which we also used in rPosNet. However, we are interested in studying purely position-based self-attention approaches and how they can perform at least on par with the (content-based) Transformer. Additionally, we compare with Shaw et al. (2018) as an upper bound since it leverages token-token interactions and was proposed for machine translation.

9 Conclusion

We have introduced the gated token-mixing approaches aPosNet and rPosNet in order to find a high-quality self-attention alternative for machine translation whose attention weights can be pre-computed at inference time. Although their token-mixing weights are position-based, the gating mechanism introduces content dependency in the form of latent weights β, as shown by our analysis. These weights capture token-token interactions and are crucial for the results of rPosNet. In our experiments, we have compared aPosNet and rPosNet with existing position-based token-mixing approaches and found that rPosNet outperforms all the position-based alternatives and performs on par with (Shaw et al., 2018) on most benchmarks while saving more than 20% of the self-attention parameters. Moreover, the possibility of pre-computing rPosNet's token-mixing weights paves the way for high-quality machine translation on specialized hardware accelerators.

Limitations

The goal of this paper is to find alternatives for self-attention with minimal or no quality loss that

can pre-compute token-mixing weights at inference time. We have compared numerous approaches across many data conditions and model sizes to show the validity of our results. However, we can identify the following limitations in our work:

- rPosNet's position-based attention is an effective replacement of Transformer's self-attention, but its usage in cross-attention leads to quality loss;
- We did not have enough computational resources to run our numerous experiments multiple times, so we relied on the consistent results we obtained across different conditions and metrics.
- While our work is motivated by future use in memristor-based devices, we have no experiments in that specific hardware because i) it is still experimental and hard to find, and ii) our proposed models still contain operations that cannot be performed naively in the analog domain.

Acknowledgments

This work was partially supported by NeuroSys which, as part of the initiative "Clusters4Future", is funded by the Federal Ministry of Education and Research BMBF (03ZU1106DA). The work reflects only the authors' views and the funding party is not responsible for any use that may be made of the information it contains.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, and Lucia Specia, editors. 2014. *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Baltimore, Maryland, USA.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Tyler Chang, Yifan Xu, Weijian Xu, and Zhuowen Tu. 2021. Convolutions and self-attention: Re-interpreting relative positions in pre-trained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4322–4333, Online. Association for Computational Linguistics.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan Loddon Yuille, and Yuyin Zhou. 2021. Transunet: Transformers make strong encoders for medical image segmentation. *ArXiv*, abs/2102.04306.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Leon Ong Chua. 1971. Memristor-the missing circuit element. *IEEE Transactions on Circuit Theory*, 18:507–519.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *CoRR*, abs/1901.02860.

Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 933–941. JMLR.org.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Linhao Dong, Shuang Xu, and Bo Xu. 2018. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Philipp Dufter, Martin Schmitt, and Hinrich Schütze. 2020. Increasing learning efficiency of self-attention networks through direct position interactions, learnable temperature, and convoluted attention. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3630–3636, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Marcello Federico, Sebastian Stüker, and François Yvon, editors. 2014. *Proceedings of the 11th International Workshop on Spoken Language Translation: Evaluation Campaign*. Lake Tahoe, California.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252. PMLR.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Dan Hendrycks and Kevin Gimpel. 2016. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, abs/1606.08415.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9:1735–80.

Miao Hu, Catherine E. Graves, Can Li, Yunning Li, Ning Ge, Eric Montgomery, Noraica Davila, Hao Jiang, R. Stanley Williams, J. Joshua Yang, Qiangfei Xia, and John Paul Strachan. 2018. Memristor-based analog computation and neural network classification with a dot product engine. *Advanced Materials*, 30(9):1705914.

Zhiheng Huang, Davis Liang, Peng Xu, and Bing Xiang. 2020. Improve transformer models with better relative position embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3327–3335, Online. Association for Computational Linguistics.

Irina Kataeva, Shigeki Ohtsuka, Hussein Nili, Hyungjin Kim, Yoshihiko Isobe, Koichi Yako, and Dmitri Strukov. 2019. Towards the development of analog neuromorphic chip prototype with 2.4m integrated memristors. In *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5.

Guolin Ke, Di He, and Tie-Yan Liu. 2020. Rethinking positional encoding in language pre-training. *CoRR*, abs/2006.15595.

Kwangyoun Kim, Felix Wu, Yifan Peng, Jing Pan, Prashant Sridhar, Kyu J. Han, and Shinji Watanabe. 2023. E-branchformer: Branchformer with enhanced merging for speech recognition. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 84–91.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. In *International Conference on Learning Representations*.

Shun Kiyono, Sosuke Kobayashi, Jun Suzuki, and Kentaro Inui. 2021. SHAPE: Shifted absolute position embedding for transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3309–3321, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. 2022. FNet: Mixing tokens with Fourier transforms. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4296–4313, Seattle, United States. Association for Computational Linguistics.

Hanxiao Liu, Zihang Dai, David So, and Quoc V Le. 2021. Pay attention to MLPs. In *Advances in Neural Information Processing Systems*.

Xuanqing Liu, Hsiang-Fu Yu, Inderjit Dhillon, and Cho-Jui Hsieh. 2020. Learning to encode position for transformer with continuous dynamical model. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6327–6335. PMLR.

Masato Neishi and Naoki Yoshinaga. 2019. On the relation between position information and sentence length in neural machine translation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 328–338, Hong Kong, China. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward

Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, New Orleans, Louisiana. Association for Computational Linguistics.

Rupesh K Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Training very deep networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Dmitri B. Strukov, Greg Snider, Duncan R. Stewart, and R. Stanley Williams. 2008. The missing memristor found. *Nature*, 453:80–83.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. 2021. Mlp-mixer: An all-mlp architecture for vision. In *Advances in Neural Information Processing Systems*, volume 34, pages 24261–24272. Curran Associates, Inc.

Zhaopeng Tu, Yang Liu, Zhengdong Lu, Xiaohua Liu, and Hang Li. 2017. Context gates for neural machine translation. *Transactions of the Association for Computational Linguistics*, 5:87–99.

Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, koray kavukcuoglu, Oriol Vinyals, and Alex Graves. 2016. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Benyou Wang, Donghao Zhao, Christina Lioma, Qiuchi Li, Peng Zhang, and Jakob Grue Simonsen. 2020. Encoding word order in complex embeddings. In *International Conference on Learning Representations*.

Zhongrui Wang, Can Li, Peng Lin, Mingyi Rao, Yongyang Nie, Wenhao Song, Qinru Qiu, Yunning Li, Peng Yan, John Paul Strachan, Ning Ge, Nathan McDonald, Qing wu, Miao Hu, Huaqiang Wu, Stan Williams, Qiangfei Xia, and Jianhua Joshua Yang. 2019. In situ training of feed-forward and recurrent convolutional memristor networks. *Nature Machine Intelligence*, 1:434–442.

Chuhan Wu, Fangzhao Wu, and Yongfeng Huang. 2021. DA-transformer: Distance-aware transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2059–2068, Online. Association for Computational Linguistics.

Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. In *International Conference on Learning Representations*.

Cheng-Xin Xue, Yen-Cheng Chiu, Ta-Wei Liu, Tsung-Yuan Huang, Je-Syu Liu, Chang Ting-Wei, Hui-Yao Kao, Jing-Hong Wang, Shih-Ying Wei, Chun-Ying Lee, Sheng-Po Huang, Je-Min Hung, Shih-Hsih Teng, Wei-Chen Wei, Yi-Ren Chen, Tzu-Hsiang Hsu, Yen-Kai Chen, Yun-Chen Lo, Tai-Hsing Wen, and Meng-Fan Chang. 2021. A cmos-integrated compute-in-memory macro based on resistive random-access memory for ai edge devices. *Nature Electronics*, 4:1–10.

Peng Yao, Huaqiang Wu, Bin Gao, Jianshi Tang, Qingtian Zhang, Wenqiang Zhang, Jianhua Joshua Yang, and He Qian. 2020. Fully hardware-implemented memristor convolutional neural network. *Nature*, 577:641–646.

Weiqiu You, Simeng Sun, and Mohit Iyyer. 2020. Hard-coded Gaussian attention for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7689–7700, Online. Association for Computational Linguistics.

A Formulas describing related position-based token-mixing approaches

In the following, we provide the formulas describing how the position-based token-mixing approaches from Section 2.2 formulate the context vector.

FNet

$$c_n := \mathcal{R}\Big(\sum_m \exp\big[-2\pi j \frac{n \cdot m}{M}\big] \cdot \mathcal{F}_h(x_m)\Big) \tag{11}$$

GaussianNet

$$c_n := \frac{1}{\sigma\sqrt{2\pi}} \sum_m \exp\Big[\frac{-(m-\mu(n))^2}{2\sigma^2}\Big] \cdot (W^V x_m) \tag{12}$$

LinearNet

$$c_n := \sum_m W_{nm} \cdot (W^V x_m) \tag{13}$$

LightConv

$$c_n := \sum_{k=0}^{2K} \frac{\exp W_k}{\sum_{k'=0}^{2K} \exp W_{k'}} \cdot \sigma_{\text{GLU}}(W^V x_{n+k-K}) \tag{14}$$

B Reformulating the gating mechanism with layer normalization

Substituting $z_m = \sigma_g(v_m)$ we rewrite the gating mechanism of Equation 4 as

$$\hat{c}_n := \Big[\sum_m \alpha_{nm} \cdot \text{Norm}(z_m)\Big] \odot g_n. \tag{15}$$

Similar to Section 3, we aim to rediscover the weighted sum over v_m. For this, we utilize the definition of layer normalization:

$$\text{Norm}(x) := a \odot [f_1(x)x - f_2(x)] + b, \tag{16}$$

Table 6: Comparing how different attention approaches leverage gating.

Model	Gating	Params	EN→DE		
			BLEU	BLEURT	COMET
Transformer	✗	66.5M	26.3	71.1	47.6
	✓	69.7M	26.6	71.6	49.6
Shaw et al. (2018)	✗	66.7M	26.3	71.4	48.6
	✓	69.9M	26.7	71.8	49.5
Rel. Self-Attention	✗	63.6M	25.7	70.4	46.4
	✓	66.7M	26.5	71.3	49.0
aPosNet	✗	61.8M	25.4	69.4	42.6
	✓	65.0M	25.9	70.6	46.1
rPosNet	✗	60.8M	25.3	70.1	45.1
	✓	63.9M	26.6	71.4	48.6

with gain $a \in \mathbb{R}^D$, bias $b \in \mathbb{R}^D$, $f_1(x) = \frac{1}{\sqrt{\sigma_x}}$ and $f_2(x) = \frac{\mu_x}{\sqrt{\sigma(x)}}$. The insertion into Equation 15 gives us:

$$\begin{aligned}\hat{c}_n \approx a \odot \sum_m \alpha_{nm} \underbrace{f_1(z_m) \cdot g_n \odot \sigma_s(v_m)}_{\beta_{nm} \in \mathbb{R}^D} \odot v_m \\ - a \odot \sum_m \alpha_{nm} \cdot f_2(z_m) \cdot g_n + \sum_m \alpha_{nm} b \odot g_n.\end{aligned} \tag{17}$$

Utilizing the normalization property $\sum_m \alpha_{nm} = 1$ we can simplify Equation 17 to:

$$\begin{aligned}\hat{c}_n \approx a \odot \sum_m \alpha_{nm}\beta_{nm} \odot v_m \\ + g_n \odot \Big[b - a \sum_m \alpha_{nm} f_2(z_m)\Big],\end{aligned} \tag{18}$$

with

$$\beta_{nm} = f_1(z_m) \cdot g_n \odot \sigma_s(1.702 v_m). \tag{19}$$

Although Equation 18 assumes α to be normalized, not normalizing α does not affect β and only adds a context-dependent scale in front of b. All in all, the Equations show that with or without layer normalization, gating introduces the token-mixing weights β.

C Theoretical complexity comparison

We compare theoretical complexities across position-based token-mixing approaches, the Transformer, and Shaw et al. (2018) concerning the number of operations and parameters in Table 7.

Table 7: We compare the theoretical complexity and number of parameters per attention layer. $\hat{K}$ refers to the bidirectional context size. With the formulation of position-based attention, the attention energies can be pre-computed after training, resulting in different complexities between training and search.

Model	Parameters		Operations	
	Train	Search	Train	Search
Transformer	$4D^2$		$2N^2D + 4ND^2$	
Shaw et al. (2018)	$\hat{K}D + 4D^2$		$\hat{K}ND + 2N^2D + 4ND^2$	
FNet	$2D^2$		$N\log(N)D + D\log(D)N$	
GaussianNet	$2D^2$		$\hat{K}ND + 2ND^2$	
LinearNet	$HN^2 + 2D^2$		$N^2D + 2ND^2$	
LightConv	$H\hat{K} + 3D^2$		$\hat{K}ND + 3ND^2$	
gLinearNet	$HN^2 + 3D^2$		$N^2D + 3ND^2$	
aPosNet	$5D^2$	$HN^2 + 3D^2$	$2N^2D + 5ND^2$	$N^2D + 3ND^2$
rPosNet	$\hat{K}D + 4D^2$	$H\hat{K}N + 3D^2$	$\hat{K}ND + N^2D + 4ND^2$	$N^2D + 3ND^2$

D Table: The impact of gating and query-key information

By depicting COMET scores in Figure 2, we visualized how the effectiveness of gating decreases with increased token-mixing weight expressiveness. In Table 6, we provide the full results with the number of parameters, BLEU, BLEURT, and COMET.

E Example failure cases of BLEU

Throughout our analysis, we observed that BLEU often disagrees with the semantic metrics BLEURT and COMET. For example, the translation quality in the Base configuration on EN→DE of GaussianNet, LinearNet, (see Table 2), aPosNet without gating, and rPosNet without gating (see Table 6) is similarly measured by BLEU but varies significantly in BLEURT and COMET. We analyzed translation samples of GaussianNet and LinearNet (see Table 8) and observed that BLEU often falsely depicts translation quality when hypotheses have little overlap with the reference or changing a single word alters the meaning of the sentence. While the inaccuracies of BLEU are already known (Kocmi et al., 2021), we want to show exemplarily how BLEU would have misled our analysis. Without using BLEURT and COMET, we would have concluded that aPosNet and rPosNet would be equally good without gating and that the hard-coded weights of GaussianNet are as good as the learnable weights of LinearNet.

Table 8: Example failure cases on EN→DE in which BLEU depicts a misleading score. These inaccurate BLEU scores are best visualized when comparing GaussianNet and LinearNet. Both models achieve the same corpus-level BLEU score but differ significantly in BLEURT and COMET (see Table 2). The translations show that measuring the syntactical overlap between the hypothesis and reference translation is not an accurate measure of translation quality.

		BLEU	BLEURT	COMET
Source	Haigerloch: Focus on the Abendmahlskirche			
Reference	Haigerloch: Abendmahlskirche rückt in den Blickpunkt			
LinearNet	Haigerloch: Fokus auf die Abendmahlskirche	15.2	84.0	72.3
GaussianNet	Haigerloch: Focus on the Abendmahlskirche	15.2	35.6	−15.0
Source	Does he know about phone hacking?			
Reference	Weiß er über das Telefon-Hacking Bescheid?			
LinearNet	Weiß er von Telefonhacking?	15.8	80.2	72.5
GaussianNet	Kennt er über Telefon-Hacking?	17.0	38.2	8.8
Source	The new season in the Falkenberg "Blue Velvet" club has begun.			
Reference	Die neue Saison in der Falkenberger Discothek "Blue Velvet" hat begonnen.			
LinearNet	Die neue Saison im Falkenberg "Blue Velvet" Club hat begonnen.	33.1	75.3	85.7
GaussianNet	Die neue Saison im Falkenberg "Blue Velvet" hat begonnen.	53.7	72.2	74.5
Source	Finally, let's talk pumpkins.			
Reference	Aber kommen wir endlich zu den Kürbissen.			
LinearNet	Abschließend möchte ich noch auf die Kürbisse eingehen.	4.8	71.4	41.0
GaussianNet	Schließlich, lassen Sie uns reden Kürbisse.	5.5	36.0	−60.3
Source	A combined English literature and language course will be scrapped.			
Reference	Der kombinierte Kurs aus englischer Literatur und Sprache wird abgeschafft.			
LinearNet	Eine kombinierte englische Literatur und Sprachkurs wird verschrottet.	9.6	60.8	44.0
GaussianNet	A combined German literature and language course will be scrapped.	3.7	19.8	−42.8
Source	However, there was no sigh of relief to be heard from Ludwigsburg.			
Reference	Ein erstes Aufatmen war aus Ludwigsburg dennoch nicht zu vernehmen.			
LinearNet	Von Ludwigsburg war jedoch kein Seufzer der Erleichterung zu hören.	5.3	76.7	46.7
GaussianNet	Es gab jedoch keinen Seufzer der Erleichterung, von Ludwigsburg gehört zu werden.	3.7	45.1	−30.3
Source	Sayings come from the Bible			
Reference	Sprichwörter kommen aus der Bibel			
LinearNet	Sprichwörter stammen aus der Bibel	42.7	90.3	108.0
GaussianNet	Sayings kommen aus der Bibel	66.9	60.7	3.7
Source	Uwe Link has an offer for anyone who wants to set off in a carriage.			
Reference	Wer dann mit der Kutsche vorfahren will, für den hat Uwe Link ein Angebot.			
LinearNet	Uwe Link hat ein Angebot für jeden, der in einer Kutsche starten will.	9.0	70.0	59.0
GaussianNet	Uwe Link hat ein Angebot für jeden, der einen Wagen starten möchte.	8.5	46.3	−10.0
Source	Solicitors should uphold the highest standards of integrity and should instil trust and confidence in the public.			
Reference	Anwälte müssen die höchsten Standards an Integrität aufrechterhalten und in der Öffentlichkeit für Vertrauen und Zuversicht sorgen.			
LinearNet	Die Staatsanwälte sollten die höchsten Standards der Integrität wahren und Vertrauen in die Öffentlichkeit schaffen.	10.9	77.9	67.4
GaussianNet	Die Umweltschützer sollten die höchsten Standards der Integrität einhalten und Vertrauen und Vertrauen in die Öffentlichkeit schaffen.	12.2	53.6	−0.7

Visual Prediction Improves Zero-Shot Cross-Modal Machine Translation

Tosho Hirasawa[†] **Emanuele Bugliarello**[⋆]
Desmond Elliott[⋆] **Mamoru Komachi**[‡]
[†]Tokyo Metropolitan University
[⋆]Department of Computer Science, University of Copenhagen
[‡]Hitotsubashi University
hirasawa-tosho@ed.tmu.ac.jp

Abstract

Multimodal machine translation (MMT) systems have been successfully developed in recent years for a few language pairs. However, training such models usually requires tuples of a source language text, target language text, and images. Obtaining these data involves expensive human annotations, making it difficult to develop models for unseen text-only language pairs. In this work, we propose the task of **zero-shot cross-modal machine translation** aiming to transfer multimodal knowledge from an existing multimodal parallel corpus into a new translation direction. We also introduce a novel MMT model with a visual prediction network to learn visual features grounded on multimodal parallel data and provide pseudo-features for text-only language pairs. With this training paradigm, our MMT model outperforms its text-only counterpart. In our extensive analyses, we show that (i) the selection of visual features is important, and (ii) training on image-aware translations and being grounded on a similar language pair are mandatory. Our code are available at https://github.com/toshohirasawa/zeroshot-crossmodal-mt

1 Introduction

Multimodal machine translation (MMT) aims to improve translation quality with the help of other modalities, such as images (Specia et al., 2016) or videos (Wang et al., 2019). MMT models have shown promising improvement over their text-only neural machine translation (MT) counterparts, especially when it matters (Li et al., 2021; Lala and Specia, 2018; Gella et al., 2019). While prior work has successfully developed MMT models for language pairs with available multimodal parallel corpora, incorporating visual information into language pairs with no multimodal dataset has received limited attention. As shown in Table 1, multimodal parallel corpora are only available for a few language pairs (Elliott et al., 2016, 2017; Barrault et al., 2018; Nakayama et al., 2020; Sanayai Meetei et al., 2019; Wang et al., 2019), which is quite less than the language pairs with text-only parallel corpora. Since building a multimodal parallel corpus by professional translators is costly and time-consuming (Wang et al., 2019), creating high-quality multimodal parallel corpora for many language pairs is not feasible.

Modality	Lang.	Examples
Text	> 700	BG, CS, DA, DE, EL, ES, ET, FR, JA, ...
Text+Image	~ 10	DE, FR, CS, JA, ...

Table 1: Number of target languages with text-only (Text) or multimodal (Text+Image) parallel corpora for the translation from English.

One approach to addressing this problem is zero-shot cross-lingual transfer, which has proven successful in text-only machine translation (Firat et al., 2016; Johnson et al., 2017, *inter-alia*). In this paper, we investigate whether this success also extends to a multimodal setting. To this end, we propose the task of **zero-shot cross-modal machine translation**, where models need to perform multimodal machine translation in language pairs that lack multimodal parallel training data. In this task, there are still language pairs with multimodal training data, but the target language pairs consist of text-only training data.

To tackle this novel task, we propose a simple **M2KT-VPN** method that aims at performing Multimodal Knowledge Transfer via Visual Prediction Network in the *zero-shot cross-modal* translation setup. Inspired by El-

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 522–535
December 6–7, 2023.

liott and Kádár (2017), a visual prediction network is employed to mimic visual features from the textual modality. We hypothesize that the predicted feature can help bridge the gap between text-only and multimodal translation pairs, so the model is not surprised when it receives true images at inference time.

The contributions of this work are as follows:

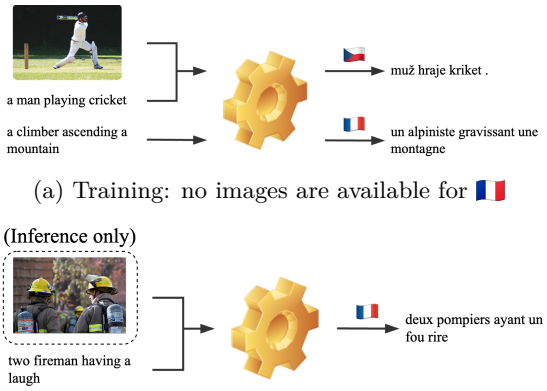
- We introduce a novel task, namely **zero-shot cross-modal machine translation** task, aiming to build MMT systems that can transfer multimodal knowledge from multimodal language pairs into text-only language pairs.
- We propose the **M2KT-VPN** model, a Transformer-based MMT model along with a visual prediction network, and show its zero-shot cross-modal translation capability.
- Our findings suggest the importance of image-aware translations and language similarity between translation directions.

2 Zero-shot Cross-Modal Machine Translation

We propose a new challenge for multimodal machine translation systems that we denote **zero-shot cross-modal machine translation** (Figure 1). This task is motivated by the real-world lack and cost of multimodal parallel corpora, which inhibits the development of multimodal translation systems beyond a few, mostly Indo-European, language pairs.

Task definition. The *zero-shot cross-modal* machine translation task aims to transfer multimodal knowledge learned from a (visually) grounded language pair into a language pair with no multimodal information at training. We define the two types of machine translation resources used for this task as follows:

- **Grounded language pairs**: language pairs where a multimodal parallel corpus is available, both at training and test time.
- **Zero-shot language pairs**: language pairs that only have a text parallel corpus for training, but have multimodal parallel data for test.



(a) Training: no images are available for 🇫🇷

(b) Inference: images are provided for 🇫🇷

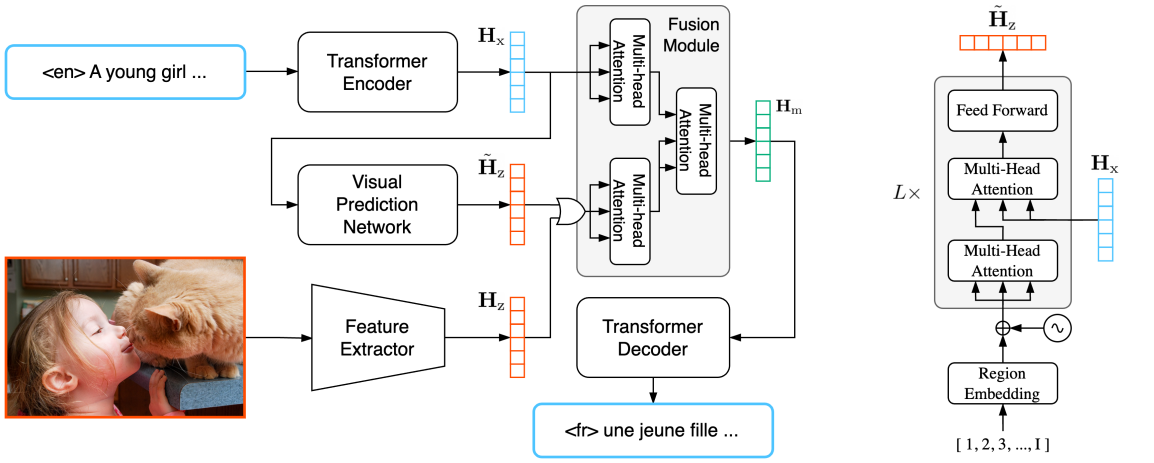
Figure 1: Overview of the **zero-shot cross-modal machine translation** task. For the zero-shot language pair (e.g., 🇫🇷), images are unavailable during training (a), but given at the inference (b).

Thus, a model is encouraged to transfer multimodal knowledge learned from grounded language pairs to zero-shot ones in order to best leverage multimodal data that may be available at test time.

Notation. We consider the following setup in our paper. Given a sequence of N tokens in a given source language, $\mathbf{x} = \langle x_1, x_2, \cdots, x_N \rangle$, and its associated image z, a multimodal machine translation model learns to translate $\mathbf{x}$ into a sentence of M tokens in a target language, $\mathbf{y} = \langle y_1, y_2, \cdots, y_M \rangle$. In the following, we directly consider a dense representation of the image z given by a visual feature extractor, which outputs I features that are then projected into a given model dimension d, $\mathbf{H}_z \in \mathbb{R}^{I \times d}$.

3 Proposed Approach: M2KT-VPN

In this section, we introduce a new MMT model, called **M2KT-VPN**, which aims to transfer multimodal knowledge learned from the multimodal corpus into the zero-shot language pair. M2KT-VPN comprises four modules (Figure 2): a Transformer (Vaswani et al., 2017) encoder to encode a source sentence, a visual prediction network (VPN) to predict a visual feature, a fusion module to incorporate multimodal information, and a Transformer decoder to generate a system output. All modules are trained simultaneously on grounded and zero-shot language pairs.



(a) M2KT-VPN (b) Visual prediction network

Figure 2: The overview of the M2KT-VPN model (a) and the visual prediction network (b).

3.1 Multilingual Machine Translation Module

We design M2KT-VPN as a multilingual MMT model. Following Fan et al. (2021), we prepend a special token (*e.g.*, `<en>`) to the source sentence $\mathbf{x}$ indicating the source language, and another special token (*e.g.*, `<fr>`) to the target sentence $\mathbf{y}$ indicating the target language. Similarly, for inference, we condition the decoder to generate a translation in a given target language by prepending its language indicator token as the first token of the sequence to be generated. We employ a cross-entropy loss to train M2KT-VPN models.

3.2 Attention-based Fusion Module

The Transformer encoder embeds a source text $\mathbf{x}$ into a high-dimensional representation $\mathbf{H}_x \in \mathbb{R}^{N\times d}$ without any presence of images. We then introduce a fusion module to ground the text-only representation $\mathbf{H}_x$ into the image z through its corresponding visual feature $\mathbf{H}_z$. This grounded representation of the source sequence $\mathbf{H}_m \in \mathbb{R}^{N\times d}$ constitutes the input to the Transformer decoder.

We use an attention-based module to fuse the visual input into multimodal representations of language. Our module first applies two dedicated self-attention operations on the text and visual features:

$$\mathbf{H}'_x = \mathrm{MHA}(\mathbf{H}_x, \mathbf{H}_x, \mathbf{H}_x) \quad (1)$$

$$\mathbf{H}'_z = \mathrm{MHA}(\mathbf{H}_z, \mathbf{H}_z, \mathbf{H}_z) \quad (2)$$

where MHA denotes the multi-head attention function (Vaswani et al., 2017). Then, a cross-attention module fuses these representations to get the multimodal representation $\mathbf{H}_m$:

$$\mathbf{H}_m = \mathrm{MHA}(\mathbf{H}'_x, \mathbf{H}'_z, \mathbf{H}'_z) \quad (3)$$

3.3 Visual Prediction Network

As described so far, our MMT model assumes the input is complete, having both text and image available for translation, both during training and inference. However, in the zero-shot cross-modal machine translation task, the visual modality is absent during training for the zero-shot language pairs.

To mitigate this gap, we propose a **Visual Prediction Network** (**VPN**) to mimic visual features for zero-shot language pairs during training. The VPN generates visual predictions from the text encoder representation $\mathbf{H}_x$. The generated visual predictions $\tilde{\mathbf{H}}_z$ in a zero-shot pair are then fed into the fusion module instead of the visual feature $\mathbf{H}_z$.

To predict the visual features corresponding to I image regions, VPN first embeds learnable visual queries (*e.g.*, Lee et al., 2018; Alayrac et al., 2022; Mañas et al., 2023; Li et al., 2023), adds positional information, and then applies layer normalization to obtain the position-aware region representations $\tilde{\mathbf{H}}^0_z$.

$$\tilde{\mathbf{H}}^0_{z,i} = \mathrm{LayerNorm}(\mathbf{E}_z(i) + \mathrm{PE}(i)) \quad (4)$$

where $\mathbf{E}_z(i)$ is the embedding representation for the i-th region, and $\mathrm{PE}(i)$ is the positional embedding for the i-th region.

The following L layers are the same as in a standard Transformer decoder, each comprising a self-attention, cross-attention, and a pairwise feed-forward module.[1] The l-th layer takes the output of the previous layer $\tilde{\mathbf{H}}_{\mathrm{z}}^{l-1}$ as input. The cross-attention module in the l-th layer takes the output of the self-attention module as the query and the text encoder output $\mathbf{H}_{\mathrm{x}}$ as the key and value. The M2KT-VPN model uses the output representation of the final layer as the visual prediction:

$$\tilde{\mathbf{H}}_{\mathrm{z}} = \tilde{\mathbf{H}}_{\mathrm{z}}^{L} \tag{5}$$

The VPN module is trained on grounded language pairs, using a max-margin loss (Elliott and Kádár, 2017) in a contrastive learning manner (Radford et al., 2021a). Given a batch of K examples, we first generate K ($\tilde{\mathbf{H}}_{\mathrm{z}}$, $\mathbf{H}_{\mathrm{z}}$) pairs. We then compute a max-margin loss for the batch:

$$\sum_{p \neq k}^{K} \sum_{i=1}^{I} \max\{0, \alpha - d({}_{k}\tilde{\mathbf{H}}_{\mathrm{z},i}, {}_{k}\mathbf{H}_{\mathrm{z},i}) + d({}_{k}\tilde{\mathbf{H}}_{\mathrm{z},i}, {}_{p}\mathbf{H}_{\mathrm{z},i})\} \tag{6}$$

where ${}_{j}\tilde{\mathbf{H}}_{\mathrm{z},i}$, ${}_{j}\mathbf{H}_{\mathrm{z},i}$ is the predicted i-th vector and the true i-th vector of j-th example in the batch; d is a cosine similarity function; and α is the margin[2]. The max-margin loss is merged with the cross-entropy loss with a coefficient of 1.0 to obtain the final loss.

4 Experiments

4.1 Experimental Setting

Dataset. We train and evaluate models on `Multi30K` dataset. We select English–Czech as a grounded language pair and English–French as a zero-shot language pair. For the training, we divide the training split of `Multi30K` into two folds of the same size; one for the grounded language pair and the other for the zero-shot language pair. The validation splits for grounded and zero-shot language pairs have the same source language texts and the target language texts, but images are absent for the zero-shot language pair. The test splits are also the same, and images are available for both grounded and zero-shot language pairs. Table 2 shows the statistics of each split. We follow a standard evaluation to report performance on four test sets: test_2016_flickr (**2016**), test_2017_flickr (**2017**), test_2017_mscoco (**mscoco**), and test_2018_flickr (**2018**).

Split	Images	Sents.
Grounded (English–Czech)		
Training	14,500	14,500
Validation	1,014	1,014
Test	2,071	2,071
Zero-shot (English–French)		
Training	–	14,500
Validation	–	1,014
Test	3,532	3,532

Table 2: The number of examples in each split for the grounded and zero-shot language pairs.

Preprocessing. For textual modality, we use Moses (Koehn et al., 2007) to lowercase, normalize punctuation, and tokenize the source and target sentences. We then learn byte pair encoding (Sennrich et al., 2016) with 10,000 merge operations on the concatenation of the training text over all language pairs to obtain a shared vocabulary for all languages. For visual modality, we extract a visual feature using **DETR-ResNet-50-DC5**[3] (Carion et al., 2020), which is an object detection model backed by a **ResNet-50** model (He et al., 2016). **DC5** stands for dilated C5 stage, which increases the feature resolution and consequently provides more information for the small objects. The extracted feature has 100 bounding boxes, each with a visual representation of 256 dimensions.

Model. We use a tiny version of the Transformer model (Transformer-tiny) as our text-only baseline and the relying model of M2KT-VPN, as this smaller model works better on `Multi30K` (Wu et al., 2021; Li et al., 2022b). This model comprises four encoder layers and four decoder layers, and the model hidden size of both decoder and decoder is 128. It also has a smaller number of attention heads and a hidden size of pair-wise feedforward network, 4 and 256, respectively. The vocabulary and

[1] We use $L = 1$ in our experiments.

[2] We use $\alpha = 0.1$ in our experiments.

[3] facebook/detr-resnet-50-dc5

embedding weights are shared across all languages. We compare our model against some baseline models:

- Transformer: a text-only Transformer-tiny model trained only on English–French data.
- mTransformer: a text-only multilingual Transformer-tiny model trained on both English–Czech and English–French data.
- IMAGINATION: a text-only multilingual Transformer with a VPN module. This model also trained on both English–Czech and English–French data.

Implementation details. We implement our models on the Fairseq (Ott et al., 2019) toolkit. The optimizer is Adam (Kingma and Ba, 2015) with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. The learning rate warms up from $1e-7$ to 0.005 over $2,000$ steps, then decays with the `inverse_sqrt` scheduler. We apply label smoothing of 0.1 for computing the cross-entropy loss and the dropout of 0.3. Early stopping with a patience of 10 is used to stop training models. We average the last ten checkpoints and use beam search with width=5 for inference.

Metrics. We train all models three times with different seeds and report averaged 4-gram BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) scores for all test sets. Additional to the classic n-gram matching evaluation, we also compute the COMET score (Rei et al., 2020)[4]. We also report statistical significance ($p < 0.05$) on the difference in BLEU scores[5].

4.2 Results

The results of our experiments are shown in Table 3. We found that our M2KT-VPN model provides an improvement over the text-only baselines and IMAGINATION model for all four test sets. The M2KT-VPN model achieves an averaged improvement of 2.65% over the mTransformer model (varies from 1.90% to 3.70% across the test sets). This performance gain would be owed to the multitask learning of the visual prediction network; the module learns to predict visual features and tailor the features for the machine translation task simultaneously.

[4]We use **Unbabel/wmt22-comet-da** (Rei et al., 2022).

[5]We used Moses' **bootstrap-hypothesis-differencesignificance.pl.**

5 Discussion

This section first provides two basic analyses of the M2KT-VPN model: model analysis and probing. We then examine various kinds of features to investigate the importance of feature selection. Finally, we ran an analysis to identify the requirement for the grounded language pair.

5.1 Model Analysis

Model ablation. Table 4 shows the results of a comprehensive ablation analysis to identify the contribution of each module in the M2KT-VPN model on entire test splits. To evaluate the contribution of the attention-based fusion module, we compare two well-known fusion strategies: concatenation-based (Li et al., 2021) and gate-based (Li et al., 2021). Firstly, the model without a VPN module drops -1.0 METEOR score, indicating a VPN module is key to resolving the missing visual modality problem in the zero-shot cross-modal machine translation task. Second, concatenation-based and gate-based models do not outperform the M2KT-VPN model and even the mTransformer baseline. The concatenation-based model fails to translate most of the examples. This evidences that attention-based fusion strategies indeed transfer multimodal knowledge.

Quality of visual prediction. Another question on M2KT-VPN is whether the visual prediction network can provide grounded visual features. To answer this question, We measured each model's Median rank score (Elliott and Kádár, 2017) on the **2016** test data. We first average true and predicted features over their regions to get every single representative vector. The predicted representative vector is compared against the true representative vectors in the test data using the cosine similarity function to produce a ranked order of the true representative vectors. The Median Rank score reports the median value of the ranks for the gold representative vector compared to the predicted representative vector.

Model	2016	2017	mscoco	2018	Average
Transformer	55.77 / 76.91	47.48 / 70.77	38.95 / 64.12	33.11 / 60.37	43.83 / 68.04
mTransformer	56.42 / 77.57	48.54 / 72.31	40.50 / 65.56	34.31 / 61.67	44.94 / 69.28
IMAGINATION	57.11 / 77.85	49.53 / 72.68	40.75 / 65.95	35.12 / 62.84	45.63 / 69.83
M2KT-VPN	**57.49** / **78.15**	†**50.19** / **73.43**	†**41.28** / **66.44**	†**35.58** / **63.06**	†**46.13** / **70.27**

Table 3: The BLEU / METEOR scores of the text-only models and MMT models in each test set for English–French translation using English–Czech as the grounded language pair. "†" indicates statistical significance of the improvement over the IMAGINATION model.

Fusion Module	VPN	BLEU	METEOR
Attention		44.79	69.27
Concatenation	✓	6.43	19.29
Gate	✓	44.88	69.42
Attention	✓	46.13	70.27

Table 4: The average BLEU and METEOR scores over all test splits for variants of M2KT-VPN.

Model	Median Rank
IMAGINATION	45.5
M2KT-VPN	47.0
Elliott and Kádár (2017)	11.0
Random	$\sim$ 500

Table 5: Median rank of randomly selected vector (Random) and model's predictions.

Model	COMET
Transformer	0.7629
mTransformer	0.7651
IMAGINATION	0.7679
M2KT-VPN	0.7698

Table 6: The averaged COMET scores over all test splits for the English–French translation.

Model	2016	2018	Average
Transformer	55.85	47.54	51.69
mTransformer	57.01	49.85	53.43
IMAGINATION	57.99	50.14	54.07
M2KT-VPN	57.78	50.79	54.28

Table 7: The METEOR scores of the text-only and MMT models in each test set for English–Czech translation

Our M2KT-VPN model returns a median rank of 47.0, which is clearly better than the random baseline. This indicates that our model is learning visually grounded representations. However, Elliott and Kádár (2017) reported a median rank of 11.0 for their RNN-based model that predicts holistic features. This difference poses another challenge to predicting region-based visual features using VPN. We would like to improve the prediction quality and explore its impact on the translation quality in our future work.

Neural-based evaluation. Table 6 shows the average COMET score over all test splits. We can see the same trend as BLEU and METEOR in Table 3. While neural-based evaluation metrics would better align with human preference than those based only on surface characteristics, this pattern may vary (Freitag et al., 2021). A human evaluation may rather be conducted to reveal which metrics align better with the text of captions, where the text is usually shorter and simpler than those in the WMT evaluation task.

Multilingualism. The multilingualism of the M2KT-VPN model is another concern. Table 7 shows the METEOR score for the English–Czech translation. The consistent improvement over the text-only baseline for both English–Czech and English–French indicates that the M2KT-VPN model is capable of performing multilingual translation.

5.2 Probing

Input degradation. We examine the model's capability of handling incomplete textual modality. Intuitively, a better MMT model can recover the content in the flawed source text from the visual modality. Following Caglayan et al. (2019) and Li et al. (2022a), we

[6] "man", "woman", "people", "mean", "girl", and "boy".

Vanilla	a	young	girl	standing	···	a	yellow	cat
Color	a	young	girl	standing	···	a	[v]	cat
Entity	a	young	girl	standing	···	a	yellow	[v]
Char.	a	young	[v]	standing	···	a	yellow	cat
Prog.	a	young	girl	standing	···	[v]	[v]	[v]

Table 8: An example of textual degradation. "Vanilla" shows the original text without degradation. "Char." and "Prog." stand for character and progressive masking, respectively. "Color" deprivation replaces words that refer to colors with a special token [v]. "Entity" and "Char." mask out the visually depictable entities and character words[6], respectively. "Prog." masking all words except the first K words. The tokens at "[v]" are masked during both training and inference.

Model	Vanilla	Color	Entity	Char.
mTransformer	77.57	71.85	61.11	70.73
M2KT-VPN	78.15 (0.03)	72.28 (-0.13)	61.49 (-0.32)	70.78 (0.04)

Table 9: The METEOR scores on vanilla, color-deprivation, entity-masking, and character-masking test sets. The scores in the parenthesis show the METEOR changes when the MMT model takes random shuffled images as its input.

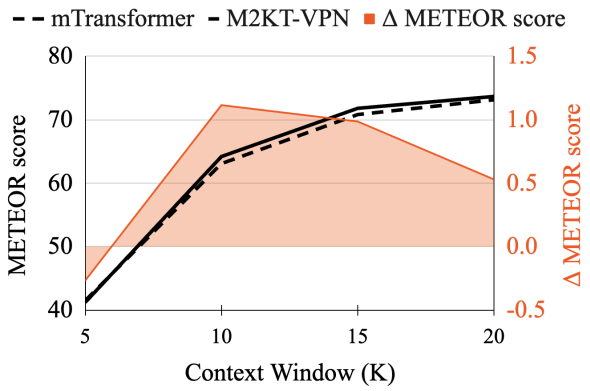


Figure 3: Evaluation with progressive masking of the context size of $\{5, 10, 15, 20\}$.

conducted four kinds of textual degradations: color deprivation, entity masking, character masking, and progressive masking. Table 8 shows examples of a complete text ("Vanilla") and its degraded ones. As entity masking is available only for the **2016** test set, we report all scores only for **2016** test set. Both the training and the test data are degraded.

Table 9 shows the BLEU and METEOR scores of the mTransformer baseline and M2KT-VPN model for vanilla, color-deprived, entity-masked, and character-masked **2016** test sets. The M2KT-VPN model outperforms the mTransformer baseline for color and entity degradation scenarios, while we see almost no change for character degradation. The possible cause of this difference is the nature of the **DETR** model we used to extract the feature. As the labels that **DETR** learns to predict contain only one word ("person") to stand for characters but more words for entities, an MMT model incorporating **DETR** would be capable of recognizing entities more precisely rather than characters. Table 10 also supports this idea. While the sentence's third [v] (corresponding to "bench") is correctly translated into "vif", the first masked entity (corresponding to "woman") keeps being mistranslated. As shown in the image, the **DETR** feature provides useful information to distinguish the "bench" from the "chair". However, it is not informative to identify the gender of the person in the image.[7]

Figure 3 compares the METEOR scores of the mTransformer baseline and an M2KT-VPN model for progressive-masked **2016** test sets with different context windows (K). The MMT model outperforms the baseline for $K = \{10, 15, 20\}$. The gap between the baseline and MMT model widens at $K = \{10, 15\}$ and narrows at $K = \{5, 20\}$. This observation for $K = \{10, 15, 20\}$ is consistent with a previous work of Li et al. (2022b), which claims the gap widens as the context window is reduced, while that for $K = 5$ is contrary to the claim. This suggests that the visual prediction network could fail to provide rich visual information when the textual context is extremely limited.

Visual awareness. We also examine the reliance of the model on the visual modality. To

[7]We found all three trained text-only systems failed to translate [v] corresponding to "bench", and all M2KT-VPN models successfully translate it.

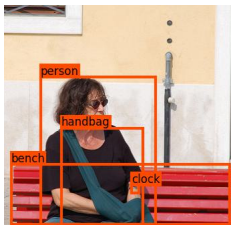
	Vanilla	the woman in the brown shirt is sitting on a bright red bench .
	Entity	the [v] in the brown [v] is sitting on a bright red [v] .
	References	la femme en t-shirt marron est assise sur un banc rouge vif .
	mTransformer	l'homme en t-shirt marron est assis sur une chaise de couleur vive . (the man in the brown t-shirt is sitting on a brightly colored chair.)
	M2KT-VPN	l'homme en t-shirt marron est assis sur un banc rouge vif . (the man in the brown t-shirt is sitting on a bright red bench.)

Table 10: Translation examples of the baseline and MMT model. The bounding boxes in the image are the prediction of the **DETR-ResNet-50-dc5** model and have a score of above 0.8. We use DeepL to translate each hypothesis into English and show it in each parenthesis.

this end, we compute the performance deterioration when a model receives incongruent images instead of congruent images (Elliott, 2018). The scores with parenthesis in Table 9 show the performance changes when the model takes incongruent images. Without surprise, the MMT model is not aware of images for vanilla, color-deprived, and character-masked test sets, as the **DETR** model does not provide rich information about color and character in an image. Meanwhile, the model is sensitive to the input image when the entities in the source text are masked out; the MMT model readily uses **DETR** feature to disambiguate the masked entities.

5.3 Visual Feature Selection

Selecting a proper visual feature has been proven to affect MMT model performance (Li et al., 2021).

In Table 11, we compare the M2KT-VPN models using different visual features extracted by different vision backbones.

- **ResNet** (He et al., 2016): An image recognition model trained to classify an image into one of the 1,000 ImageNet classes. **ResNet-50** and **ResNet-101** comprise 50 and 101 layers, respectively. We extract the local features of each ResNet model and feed them into the MMT models.
- **Faster R-CNN** (Anderson et al., 2018): An object detection model trained to segment an image into 36 salient image regions and predict the object in each region.
- **DETR** (Carion et al., 2020): A transformer-based object detection model trained to segment an image into 100 regions and predict the object in each region. We used four different backbones: **ResNet-50**, **ResNet-50-DC5**, **ResNet-101**, and **ResNet-101-DC5**.
- **CLIP** (Radford et al., 2021b): A vision and language model trained on various image and text pairs in a self-supervised way. We examined three CLIP models using different backbones: **ResNet-101**, **ViT-B/16**, and **ViT-B/32**. We use the visual encoder of each CLIP model to encode images; no textual modality is involved in the extraction process.

10 out of 11 MMT models outperform the mTransformer model in both BLEU and METEOR scores. This shows that M2KT-VPN models are capable of incorporating various kinds of visual features. The only feature that deteriorates the model performance is ResNet-101; the feature extracted by ResNet-101 would be highly optimized for image classification and not suitable for machine translation.

Among all features, **DETR** with the **ResNet-50-DC5** backbone serves as the best feature extractor for the M2KT-VPN model. On the other hand, the model using **CLIP** features obtains almost equal performance to those using **ResNet** features. This observation is partially contrary to the previous works claiming that enhanced vision features obtain superior performance compared with low-level vision features (Li et al., 2022a).

We also observed that **DETR** with **DC5** backbone outperforms the non-DC5 counterparts. As **DC5** models provide the feature with higher resolution, the MMT model can receive richer information about small objects in an image. Consequently, the MMT model can better understand and translate those small objects more accurately.

Feature	BLEU	METEOR
None (mTransformer)	44.94	69.28
ResNet-50	45.34	69.67
ResNet-101	44.79	69.29
Faster R-CNN	45.72	69.65
DETR (ResNet-50)	45.79	70.01
DETR (ResNet-50-DC5)	**46.13**	**70.27**
DETR (ResNet-101)	45.49	69.84
DETR (ResNet-101-DC5)	45.81	69.91
CLIP (ResNet-101)	45.19	69.47
CLIP (ViT-B/16)	45.64	69.88
CLIP (ViT-B/32)	45.36	69.64

Table 11: The averaged BLEU and METEOR scores over all test splits of M2KT-VPN models using different visual features. The models in the parentheses are backbone models.

Grounded	BLEU	METEOR
→ Czech	46.14 (↑1.12)	70.27 (↑0.93)
→ German	45.95 (↓2.04)	69.95 (↓1.16)
→ Japanese	42.34 (↓2.53)	68.25 (↓0.99)

Table 12: The scores over all test splits of the M2KT-VPN model using different grounded language pairs. Each "→ X" stands for English → X as the grounded language pair. The scores in parenthesis are the changes from the text-only counterpart.

5.4 Grounded Language Pairs

The ability of a model to transfer multimodal knowledge between grounded and zero-shot language pairs is another key research question for this task. To answer this question, we compare three grounded language pairs for English–French zero-shot cross-modal translation.[8]

Shown in Table 12, the translation performance of using English–Czech as a grounded language pair is better than those of using English–German and English–Japanese.

The observation of using English–German contradicts our intuition that the more similar two language pairs are, the better one serves as a grounded language pair for another. As English–German training data is generated with no involvement of images, this indicates that M2KT-VPN requires image-aware training data to transfer multimodal knowledge.

English–Japanese also contains visual-aware translations, but it does not improve the performance of English–French. We found that M2KT-VPN translated the 1.43% of entire test examples into Japanese regardless the decoder is conditioned to generate French translation[9]. This ratio is much higher than that of the text-only counterpart (0.27%) and M2KT-VPN using English–Czech (0.26%) or English–German (0.28%). We conclude that grounded and zero-shot pairs should not be too distant.

6 Related Work

Multimodal machine translation. This task has been developed along with the creation of multimodal parallel corpora. After the first multimodal parallel corpus, namely Multi30K for English–German translation, emerged at the first conference of machine translation (Bojar et al., 2016), many publicly available datasets have been proposed: the English–French version of Multi30K and new test sets at 2017 (Elliott et al., 2017), the English–Czech version of Multi30k (Barrault et al., 2018), and the English–Japanese version of Multi30k (Nakayama et al., 2020). More recently, Guo et al. (2022) proposed a private expansion of Multi30K, including Hindi, Turkish, and Latvian translations. They examined a multilingual MMT model on their dataset and investigated the multilingual ability of the model. We put the step forward and investigate the zero-shot cross-modal translation capability in an MMT task.

Predicting a visual feature from textual modality is a well-established approach for improving multimodal machine translation systems. Elliott and Kádár (2017) first divided the multimodal machine translation task into two subtasks: translation task and visual grounding task. Similarly, Zhou et al. (2018) employed a latent space learning task as their visual grounding task to bridge textual and visual modalities. Recently, Li et al. (2022b) proposed to utilize the feature prediction from a visual prediction network. We make use of the model for the visually grounding task and propose to incorporate the prediction as a pseudo-visual feature with MMT models.

[8]We retrieved Japanese translations from Flickr30kEnt-JP (Nakayama et al., 2020)

[9]We used Google's language-detection library.

Zero-shot cross-lingual machine translation. Zero-shot *cross-lingual* machine translation aims to perform a translation with zero-resource where the considering language pairs do not have any parallel corpora (Firat et al., 2016; Johnson et al., 2017; Chen et al., 2017; Lample et al., 2018; Artetxe et al., 2019). The previous works have proved the zero-shot cross-lingual translation capability.

In a multimodal setting, we are only aware of two previous efforts on zero-shot transfer. Huang et al. (2020) simulated that no parallel corpus exists between the language pair and proposed utilizing the image as the pivot and performing a zero-shot cross-lingual translation. Besides, Long et al. (2021) trained a generative adversarial network (GAN) (Goodfellow et al., 2014) for generating the visual features for text-only language pairs. Both approaches use images for training, and evaluate models on a single text-only translation direction. Unlike these works, our work (i) tests MMT models with complete multimodal inputs and (ii) takes advantage of a multilingual model.

7 Conclusion

In this paper, we proposed a new task, **zero-shot cross-modal machine translation**, aiming to evaluate MMT systems from the perspective of the cross-lingual transferability of multimodal knowledge learned from grounded language pairs into language pairs with only text data during training.

Our proposed MMT model shows promising results, suggesting that the VPN mitigates the modality mismatch between training and inference steps for zero-shot language pairs. The analysis shows the importance of selecting a proper visual feature and the necessity of image-aware translations, both of which should be key properties of MMT models.

Limitations

Although our M2KT-VPN model has shown the zero-shot cross-modal translation capability, some limitations exist. While the well-established visual features are informative for some object entities, they do not benefit the translation of character and color words. Besides, the importance of language similarity between grounded and zero-shot pairs limits the language pairs we can apply M2KT-VPN for. In future work, we will extend our M2KT-VPN model to relax this limitation.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 194–203, Florence, Italy. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4159–4170, Minneapolis, Minnesota. Association for Computational Linguistics.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, page 213–229, Berlin, Heidelberg. Springer-Verlag.

Yun Chen, Yang Liu, Yong Cheng, and Victor O.K. Li. 2017. A teacher-student framework for zero-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1925–1935, Vancouver, Canada. Association for Computational Linguistics.

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978, Brussels, Belgium. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation*, pages 215–233, Copenhagen, Denmark. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Desmond Elliott and Ákos Kádár. 2017. Imagination improves multimodal translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 130–141, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond English-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Spandana Gella, Desmond Elliott, and Frank Keller. 2019. Cross-lingual visual verb sense disambiguation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1998–2004, Minneapolis, Minnesota. Association for Computational Linguistics.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Hongcheng Guo, Jiaheng Liu, Haoyang Huang, Jian Yang, Zhoujun Li, Dongdong Zhang, and Zheng Cui. 2022. LVP-M3: Language-aware visual prompt for multilingual multimodal machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2862–2872, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Po-Yao Huang, Junjie Hu, Xiaojun Chang, and Alexander Hauptmann. 2020. Unsupervised multimodal neural machine translation with pseudo visual pivoting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8226–8237, Online. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Chiraag Lala and Lucia Specia. 2018. Multimodal lexical translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. 2018. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*.

Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and JingBo Zhu. 2022a. On vision features in multimodal machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6327–6337, Dublin, Ireland. Association for Computational Linguistics.

Jiaoda Li, Duygu Ataman, and Rico Sennrich. 2021. Vision matters when it should: Sanity checking multimodal machine translation models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8556–8562, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv*, abs/2301.12597.

Yi Li, Rameswar Panda, Yoon Kim, Chun-Fu (Richard) Chen, Rogerio S. Feris, David Cox, and Nuno Vasconcelos. 2022b. Valhalla: Visual hallucination for machine translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5216–5226.

Quanyu Long, Mingxuan Wang, and Lei Li. 2021. Generative imagination elevates machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5738–5748, Online. Association for Computational Linguistics.

Oscar Mañas, Pau Rodriguez Lopez, Saba Ahmadi, Aida Nematzadeh, Yash Goyal, and Aishwarya Agrawal. 2023. MAPL: Parameter-efficient adaptation of unimodal pre-trained models for vision-language few-shot prompting. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2523–2548, Dubrovnik, Croatia. Association for Computational Linguistics.

Hideki Nakayama, Akihiro Tamura, and Takashi Ninomiya. 2020. A visually-grounded parallel corpus with phrase-to-region linking. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4204–4210, Marseille, France. European Language Resources Association.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021a. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021b. Learning transferable visual

models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Loitongbam Sanayai Meetei, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2019. WAT2019: English-Hindi translation on Hindi visual genome dataset. In *Proceedings of the 6th Workshop on Asian Translation*, pages 181–188, Hong Kong, China. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

X. Wang, J. Wu, J. Chen, L. Li, Y. Wang, and W. Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4580–4590, Los Alamitos, CA, USA. IEEE Computer Society.

Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 2021. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6153–6166, Online. Association for Computational Linguistics.

Mingyang Zhou, Runxiang Cheng, Yong Jae Lee, and Zhou Yu. 2018. A visual attention grounding neural model for multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3643–3653, Brussels, Belgium. Association for Computational Linguistics.

A Translation Examples

Table 13 shows the translation examples for the vanilla source text.

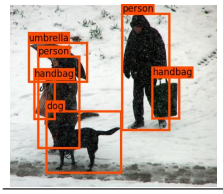
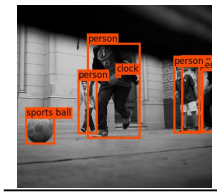
	Source	two people are walking the dog through the snow .
	Reference	deux personnes promènent leur chien dans la neige .
	mTransformer	deux personnes marchent ϕ dans la neige . (two people walking ϕ in the snow .)
	M2KT-VPN	deux personnes promènent le chien dans la neige . (two people walking the dog in the snow .)
	Source	several children are watching someone chase a ball on the sidewalk .
	Reference	plusieurs enfants regardent quelqu'un courir après une balle sur le trottoir .
	mTransformer	plusieurs enfants regardent quelqu'un ϕ sur le trottoir . (several children look at someone ϕ on the sidewalk .)
	M2KT-VPN	plusieurs enfants regardent quelqu'un après une balle sur le trottoir . (several children look at someone after a ball on the sidewalk .)

Table 13: Translation examples of the baseline and M2KT-VPN model for the vanilla source text. The bounding boxes in the image are the prediction of the **DETR-ResNet-50-dc5** model and have a score of above 0.8. We use DeepL to translate each hypothesis into English and show it in each parenthesis. The "ϕ" stands for the omitted target word in the translation.

The GENDER-GAP Pipeline: A Gender-Aware Polyglot Pipeline for Gender Characterisation in 55 Languages

Benjamin Muller, Belen Alastruey, Prangthip Hansanti, Elahe Kalbassi, Christophe Ropers, Eric Michael Smith, Adina Williams, Luke Zettlemoyer, Pierre Andrews* and Marta R. Costa-jussà*
FAIR, Meta
{benjaminmuller,alastruey,prangthiphansanti,ekalbassi,chrisopers
ems,adinawilliams,lsz,mortimer,costajussa}@meta.com

Abstract

Gender biases in language generation systems are challenging to mitigate. One possible source for these biases is gender representation disparities in the training and evaluation data. Despite recent progress in documenting this problem and many attempts at mitigating it, we still lack shared methodology and tooling to report gender representation in large datasets. Such quantitative reporting will enable further mitigation, e.g., via data augmentation. This paper describes the GENDER-GAP Pipeline (for **G**ender-**A**ware **P**olyglot Pipeline), an automatic pipeline to characterize gender representation in large-scale datasets for 55 languages. The pipeline uses a multilingual lexicon of gendered person-nouns to quantify the gender representation in text. We showcase it to report gender representation in WMT[1] training data and development data for the News task, confirming that current data is skewed towards masculine representation. Having unbalanced datasets may indirectly optimize our systems towards outperforming one gender over the others. We suggest introducing our gender quantification pipeline in current datasets and, ideally, modifying them toward a balanced representation.[2]

1 Introduction

Despite their widespread adoption, Natural Language Processing (NLP) systems are typically trained on data with social and demographic biases. Such biases inevitably propagate to our models and their generated outputs, e.g., by over-representing a given demographic group and under-representing others. It is, therefore, critical to measure, report, and design methods to mitigate these biases, before they can be encoded and potentially amplified during training (Foulds et al., 2020; Wang and Russakovsky, 2021).

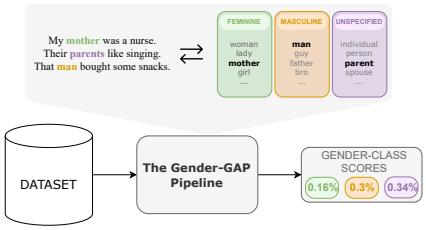


Figure 1: The Gender-GAP Pipeline works by identifying gendered lexical terms and reporting statistics on these lexical matching.

This paper focuses on quantifying gender representation in highly multilingual data (see Figure 1), in particular, for the task of machine translation. Gender is a complex concept that can be defined in many ways depending on the field of study, language or culture (Chandra et al., 1981; Hellinger and Bussmann, 2001; Kramer, 2020). We discuss and define gender in Section 3.1. However, briefly, we define gender bias as the systematic unequal treatment based on one's gender (Blodgett et al., 2020; Stanczak and Augenstein, 2021). Gender bias, when it impacts training data, may decrease the performance of the system on certain gender groups (Hovy et al., 2020). When impacting evaluation data, it may push the system designers to deploy a system that causes harm by favoring one group over others (Mehrabi et al., 2021). For example, a system that translates text that includes feminine nouns more poorly than text with masculine nouns may lead the end users to miss important information or misunderstand the sentence (Savoldi et al., 2021). A system that inaccurately translates a gender-neutral sentence in English e.g. *they are professors* to a sentence with a masculine noun *ils sont professeurs* in French may also lead to serious representational harm.

We propose the GENDER-GAP pipeline to quantify gender representation bias of multilingual texts

[1] http://www2.statmt.org/wmt23/

[2] The GENDER-GAP pipeline is available at https://github.com/facebookresearch/ResponsibleNLP/tree/main/gender_gap_pipeline

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 536–550
December 6–7, 2023.

using lexical matching as a proxy. Our pipeline can be seen as two main modules.

First, we build a multilingual gender lexicon: starting from a list of about 30 English nouns extracted from the HolisticBias dataset (Smith et al., 2022), split into 3 gendered classes—masculine, feminine, and unspecified. We manually translate them and reassign them to the appropriate gender class for each target language (e.g. "grandfathers", masculine in English, becomes "abuelos", masculine and unspecified in Spanish). Our list is restricted to nouns that refer to people (e.g. man, woman, individual) or to kinship relationships (e.g. dad, mom, parent). Most languages, including genderless languages (Prewitt-Freilino et al., 2012) (e.g. Finnish, Turkish) encode genders through kinship relationships and person terms (Savoldi et al., 2021). For this reason, focusing on a restricted list of kinship and person nouns allow us to scale our lexicon to 55 languages.

Second, we arrive at a straightforward and easily comparable gender distribution by using a word matching counter. Based on our newly collected multilingual lexicon, our pipeline segments each input sentence at the word-level using Stanza (Qi et al., 2020), a state-of-the-art word segmentation tool, and counts the number of occurrences of words in each gender class. As a result, we obtain a gender distribution across 55 languages. In summary, our contribution is threefold:

- We collect and release an aligned multilingual lexicon that can support measurement of the representation of genders in 55 languages.
- We introduce the Gender-Aware Polyglot pipeline (GENDER-GAP), a lexical matching pipeline, and describe the gender distribution observed in popular machine translation training and evaluation data. On average, all three analyzed datasets are biased toward the masculine gender. We find the gender representations to be domain- and language-specific. Additionally, using the GENDER-GAP pipeline, we can discover sentences that have been translated with a gender bias.
- We release our pipeline and recommend the reporting of gender representations in machine translation training and evaluation datasets to improve awareness on potential gender biases.

2 Related work

The study of biases in text has become more important in recent years, with Large Language Models (LLMs) displaying bias against people depending on their demographics and identity. As a testament to the importance of this topic, many recent papers, including those introducing GPT-3 and 4 (Brown et al., 2020; OpenAI, 2023), PaLM 1 and 2 (Chowdhery et al., 2022; Anil et al., 2023), LLaMa 1 and 2 (Touvron et al., 2023a,b), analyze how such biases affect their model outputs. Some works even discuss frequencies of gendered terms in their pretraining corpora (Anil et al., 2023; Touvron et al., 2023b), as this can affect downstream generation. Despite this acknowledgment of the issue, general purpose tools to measure demographic biases are still fairly rare, and so far have mainly been in English.

However, some have begun to measure demographic biases beyond English. Smith et al. (2022) built a comprehensive analysis dataset covering 13 demographic groups and Costa-jussà et al. (2023) extended it to the multilingual setting. Specific to Machine Translation, Savoldi et al. (2021) discussed best practices in reporting gender bias. Several works (Stanovsky et al., 2019; Prates et al., 2020; Renduchintala et al., 2021; Renduchintala and Williams, 2022) have explored metrics for exposing failures in automatically translating pronoun and occupations, and some have even explored MT model training (Escudé Font and Costa-jussà, 2019; Stafanovičs et al., 2020) or fine-tuning (Saunders et al., 2020; Corral and Saralegi, 2022; Costa-jussà and de Jorge, 2020) or both (Choubey et al., 2021) to lessen the effect of gender-related biases. More than this, there are initiatives that provide toolkits to generate multilingual balanced datasets in terms of gender (Costa-jussà et al., 2019) from Wikipedia and even balanced in gender within occupations (Costa-jussà et al., 2022).

However, despite the progress made, most of these resources only cover a handful of languages—the community still lacks easy to use, open-source toolkits to measure biases across a large number of languages. In this work, we address this need by showcasing, GENDER-GAP, a lexical matching pipeline to measure gender distribution across 55 languages.

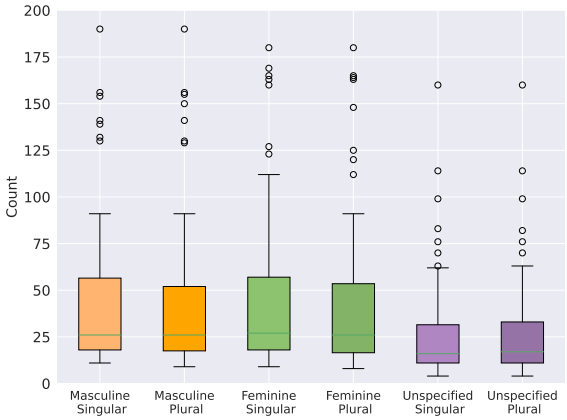


Figure 2: Distribution of the number of words in our proposed multilingual gender lexicon per language across gender-classes and number (i.e. singular and plural)

3 Proposed Data Collection and Pipeline

3.1 Defining Gender

Gender is a complex topic that can be defined in many different ways depending on the field of studies and the context (Hellinger and Bussmann, 2001). In this work, we approach gender from two perspectives:

First, linguistic gender (Corbett, 2013; Cao and Daumé III, 2020; Kramer, 2020; Stanczak and Augenstein, 2021) corresponds to the classification of linguistic units, such as words, into categories based on the gender information they provide. Linguistic gender refers to overlapping notions, such as *grammatical*, and *semantic* gender, depending on the properties of the language. Grammatical gender implies the classification of nouns, adjectives, and other parts of speech into categories based on their morphosyntactic properties. In many languages, grammatical gender morphology appears on all nouns, regardless of whether they refer to persons, animals, plants, or inanimate objects (e.g., "il libro" *the book* is a masculine noun in Italian). Semantic gender (Corbett, 1991) refers to the existence of lexical units whose meaning is associated with a specific cultural notion of peoples' gender(s). For instance, in English, the word "men" associated with masculine traits, "woman" with feminine ones, etc. Semantic gender then may be present in languages that do not morphologically mark grammatical gender, such as English, Turkish, or Mandarin Chinese. In languages that do mark grammatical gender, grammatical and semantic gender do not always match: for example, in German, the word for girl "Mädchen" is grammatically neuter, but refers to a person which would fall into our 'feminine' class based on its meaning. For our purposes, we use semantic gender classes in our multilingual lexicon, since we are interested in gender representation.

Our goal is to build and foster inclusive NLP technologies that do not carry, replicate, or amplify social gender biases, which can impact end users and societies negatively by affecting representations of specific groups. However, there are social meanings of gender that are not readily accessible in text, so, we use semantic gender on human words as a proxy for social gender.

Social gender refers to gender as a social construct based on cultural norms and identity (Ackerman 2019; Cao and Daumé III, 2020; Stanczak and Augenstein, 2021; Duignan, 2023). As highlighted by Ackerman 2019, social gender is defined as the internal gender experienced by a given human individual. For this reason, data-driven analysis of genders in large corpora can only relate to social gender indirectly through linguistic notions of gender(s).[3] We assume for our purposes that a list of gendered words can be used to approximate some important aspects of social gender for the purposes of measuring representation disparities.

3.2 Aligned Gendered Multilingual Lexicon

To measure gender distribution across 55 languages, we first build a multilingual lexicon. We want this lexicon to be as aligned as possible across languages while also encoding language-specific gender linguistic phenomena.

Languages Our lexicon is available in 55 typologically and phylogenetically diverse languages such as English, Finnish, Zulu, Vietnamese, Ganda, Japanese or Lithuanian, spanning 15 distinct scripts. We report the complete list of languages in Figure 6.

Gender Classes We define three semantic gender classes: masculine, feminine and unspecified. The unspecified class aggregates nouns of different sorts. It mainly capture nouns that do not explicitly encode any particular gender (e.g. "person" is considered unspecified in English). For this reason,

[3]We recall that gender is distinct from sex which refers to collections of biological properties of individuals such as genes (e.g., chromosomes), phenotypes (e.g., anatomy) (Council of Europe, 2023). See Butler (2011) for a discussion of additional factors that complicate this view.

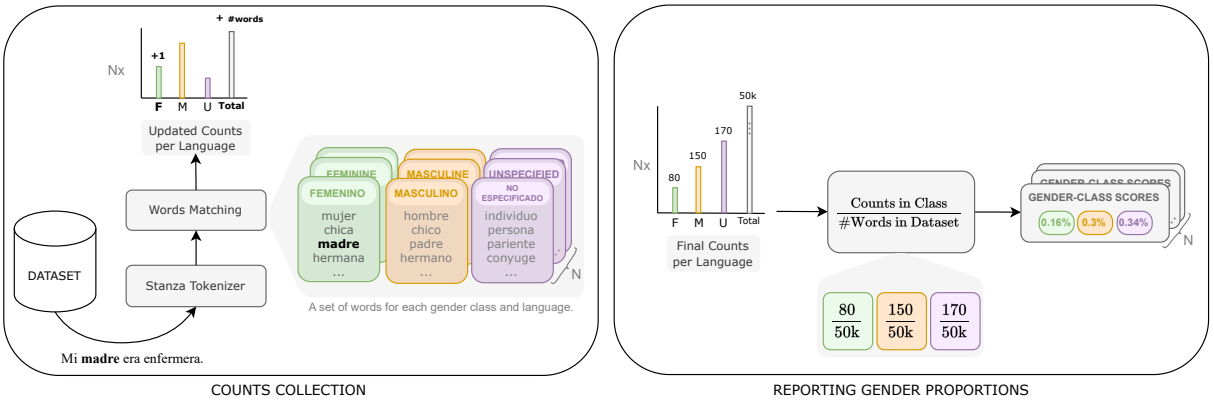


Figure 3: Diagram of the GENDER-GAP pipeline. In the first stage, we process each sentence of the 55 supported languages of the dataset and count the word matches for each category. Once this step is completed, we compute a gender-class score which corresponds to the proportion of gendered noun matched within all the words in the dataset.

"unspecified" can be seen as aggregating masculine, feminine and non-binary genders (Herdt, 2020).

While there exist more complex gender lexica as discussed in Stanczak and Augenstein (2021), they are focused on English and are not always easily translated. Because our goal is to provide a methodology that can be used to evaluate bias across multiple languages, we take a more pared down lexical approach.

Lexicon creation We start by defining a list of about ten, high frequency person nouns per gender class in English. Each noun is found in both its singular and plural form. To find a list of nouns that is as universal as possible, we restrict this list of persons such as masculine "man", feminine "woman", and "person" and synonyms (e.g. "individual") that we complement with kinship terms classified by gender (e.g., masculine "father", feminine "mother", neutral "parent"). Our list corresponds to the one defined in the previous work of HolisticBias (Smith et al., 2022), which is only available in English.[4]

We then translate these nouns into the other languages by reassigning them to the appropriate gender class. A noun in a given gender class may be part of another class (or multiple other classes) in another language. For instance "grandparents" (masculine, plural) becomes "abuelos" in Spanish which is both masculine and unspecified genders.

The English-language source list is passed on translators who are native speakers of the target language, with language proficiency at CEFR[5] level C2 in the source language. For all languages, translators are asked to provide equivalent singular and plural terms in their respective native language, except if any of the source concepts do not exist in the language. For example, not all languages use a distinctive, gender-agnostic term such as the English term *sibling*, distinctively from either *brother* or *sister*. We also consider that the reverse can be true (i.e. that the target language may have more than one term to translate one of the English terms in the source list), and give the translators the possibility to provide additional translations in such cases. For instance, when we translate *women* into Korean we get : "여성들" and "여인들".

Additionally, translators are asked to consider the terms in the source list as lemmas (or headwords in dictionary entries) and, if applicable to the given language, to provide relevant morphologically derived forms, including cases and gendered forms. Finally, translators are also encouraged to provide terms covering all language registers, which is necessary because some languages (e.g., Thai or Korean, among others) use several different terms at various levels of formality.

We are cognizant of the fact that this approach presents several limitations. The first limitation occurs when a term could be said to fall into both the unspecified and one of the gendered categories. For example, the term Spanish *padres* can be used to mean both *fathers* or *parents*. Some speakers also use the singular form to mean *parent* (and not necessarily *father*). The second limitation applies to

[4]We use the gender noun list v1.1 from HolisticBias

[5]https://coe.int/en/web/common-european-framework-reference-languages/level-descriptions retrieved 2023-07-24

languages that are closer to the synthetic end of the analytic-synthetic spectrum; i.e. languages that are agglutinative or highly fusional (e.g., Zulu, Uzbek, Estonian). This approach may not allow for the detection of many agglutinated or fused word forms. Finally, due to the templated, context-free nature of the lexicon, one term was particularly difficult to disambiguate: *veteran*, which can be used to refer to a soldier or a seasoned professional.[6] Cultural differences also had to be considered in addition to the above ambiguity; for example, Japanese translators mentioned the fact that the Japanese equivalent of the term was infrequently used with the first meaning cited above.[7]

Lexicon statistics In Figure 2 we can see the obtained data distribution across number and gender for the different languages. We notice a few outliers. As described above, translators are asked to provide relevant morphologically derived forms. This makes the number of nouns in Estonian to be 7 times larger than the average. For instance, "woman" is translated into *naine* "a woman", *naise* "of a woman", *naisele* "to a woman", etc.

3.3 Proposed Pipeline

Figure 3 shows a diagram of the GENDER-GAP pipeline. In the first stage or the counts collection, we work at the sentence level for the NTREX and FLORES-200 and at the document level for Common Crawl. We segment each sample at the word level using `Stanza` tokenizer available in the given language (Qi et al., 2020) except for Cantonese (yue) for which we reuse the model available for simplified Chinese (zh-hans) and Thai for which we use `PyThaiNLP`.[8] For the rest of the languages we use simple `nltk`[9] typographic tokenizer (based on white-space and punctuation marks). We then count and increment a gender-class counter anytime we match a word in the list of words representative of this class. For instance, in the sentence "my mother was a nurse" the pipeline will add +1 to the feminine counter (due to lexical match of "mother").

Once this process has been done for each sentence in the dataset we move to the second stage

Figure 4: Gender Representation in % of the total tokens in the FLORES dataset dev split.

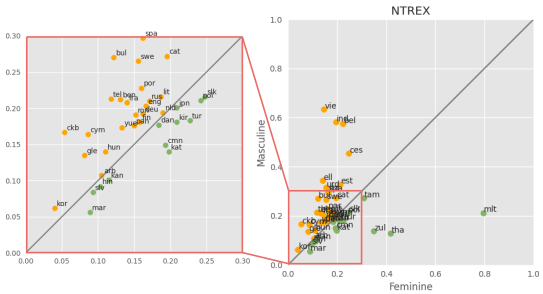


Figure 5: Gender Representation in % of the total tokens in the NTREX dataset.

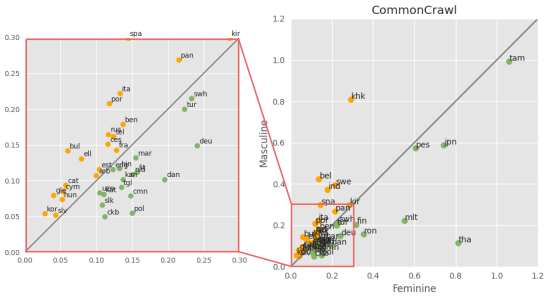


Figure 6: Gender Representation in % of the total tokens in the Common Crawl dataset.

or the reporting of gender proportions where we define a score for each gender-class by dividing the gender-class count by the total number of words in the dataset. By doing so, the final gender score does not depend on any defined linguistic macro-unit such as sentences or documents lengths but only on the word-level tokenization.

4 Experiments

To showcase GENDER-GAP, we run it on Common Crawl raw data and two popular machine translation evaluation datasets: FLORES-200 (NLLB Team et al., 2022) and NTREX-128 (Federmann et al., 2022). FLORES is a Wikipedia-based dataset including 3001 sentences translated from English to 200 languages. NTREX-128 is made of 1997

[6] https://www.merriam-webster.com/dictionary/veteran retrieved 2023-07-24

[7] See https://en.wikipedia.org/wiki/Article_9_of_the_Japanese_Constitution retrieved 2023-07-24

[8] https://pythainlp.github.io/docs/2.0/api/tokenize.html

[9] https://www.nltk.org/api/nltk.tokenize.html

Lang	Fem.	Masc.	Uns.	Δ(|Fem.-Masc.|)	%doc.
			Flores DevTest.		
eng	0.121	0.065	**0.379**	0.056 (0.0003)	11.2
avg.	0.128	0.144	**0.302**	0.097 (0.0003)	10.1
			NTREX		
eng	0.166	0.203	**0.379**	0.037 (0.0003)	15.5
avg.	0.180	0.224	**0.329**	0.099 (0.0003)	13.4
			CommonCrawl		
eng	0.120	0.115	**0.243**	0.005 (0.0000)	9.4
avg.	0.212	**0.260**	0.251	0.136 (0.0003)	12.0

Table 1: % Gender Distribution in WMT Evaluation dataset. We report the English distribution and the average across all languages (standard deviation indicated between parenthesis). The full table is available in the appendix Table 3-5. We **bold** the most represented gender class, and underline the second most represented gender class. We define the the gender gap Δ defined as the absolute difference between the Feminine and Masculine scores. %doc. refers to Coverage.

sentences from news documents originally collected for WMT 2019 (Barrault et al., 2019) translated from English into 128 languages. Both these datasets are part of the corpora provided by the WMT shared task. In addition, we run the pipeline on a sample of Common Crawl.[10] Common Crawl is a snapshot of crawlable web data that is widely used in the NLP community thanks to the release of the CCNET corpora (Wenzek et al., 2020), the OSCAR corpus (Ortiz Suárez et al., 2019) and the C4 corpus (Raffel et al., 2019). It is used to train NLP systems like language and machine translation models. We run our pipeline on 100k documents for each language. Our pipeline supports 55 languages, and we run it on the intersection of these datasets with the set of supported languages.

5 Analysis

5.1 Quantitative Analysis

We report the average coverage and gender distribution in Table 1 along with the complete tables for the 55 languages in Table 3-5.

Coverage We first look at the number of samples for which at least one noun is found (cf. %doc in Table 1). We find that, on average, about 10% of samples match with at least a noun (between 10.1 and 13.4% depending on the dataset). We find that the coverage is the largest for Vietnamese (with up 45.7% of samples matched) and Thai (28.9% of samples matched) and the smallest for Korean (between 1.7% and 2.5% depending on the dataset). This shows that even though our lexicon is restricted to person nouns and kinship relationships, we are still covering a very large number of samples based on which we measure gender representations.

Gender Distribution Table 1 shows gender representation for masculine, feminine and unspecified. For better visualization, Figures 4, 5 and 6 report the % of masculine and feminine representation of the total tokens in FLORES, NTREX, and Common Crawl respectively.

On average, the masculine gender is more represented than the feminine in all three datasets. We find that NTREX is the dataset with the highest bias toward the masculine gender on average. Accounting for uncertainty, using the standard error to define a confidence interval,[11] we find that 30/45 languages are biased toward the masculine gender for NTREX. This includes languages like English, Arabic, French, Spanish, Vietnamese, and Panjabi. The rest of the languages are either balanced between masculine and feminine (i.e. Δ(|Fem.-Masc.|) is inferior to the confidence interval length) or biased toward the feminine gender. In addition, we find 16/54 languages biased toward the masculine gender for all three datasets suggesting an inherent gender bias in these languages. This includes several romance languages such as Spanish, French, Catalan and Italian along with Belarusian, Indonesian, and Panjabi.

Impact of Domains We find that 14/55 languages for which, the gender representation changes drastically across the different datasets. For instance, the gender differences are much larger in NTREX than in Common Crawl data. More specifically, in Lithuanian the distribution is skewed toward the masculine class for NTREX data, while it is skewed toward the feminine for Common Crawl data. For Danish, the gender representation is balanced for NTREX but skewed toward the Feminine class for Common Crawl data. This shows that domains highly impact gender representation. NTREX is based on news data, while

[10] https://commoncrawl.org/

[11] We consider that a given dataset in a language is biased toward a specific gender when the gap Δ(|Fem.-Masc.|) is higher than two times the standard error (ste.). This is equivalent to defining a confidence interval as $[r_g - 2ste, r_g + 2ste]$ given the gender score r_g with $g \in \{masc., fem.\}$. If Δ(|Fem.-Masc.|) is inferior to $2ste$, we consider the dataset to be gender balanced. ste is defined as $\frac{\sigma(fem-masc)}{\sqrt{n}}$ with n the number of words in the dataset and σ the standard deviation. See (Bulmer, 1979) for more details on these definitions.

Sentence 1: Omission of words/lexical variation	
Eng: shark injures 13-year-old on lobster dive in california	`masc.+= 0`
Spa: tiburón hiere a un **niño** de 13 años que buceaba en busca de langostas en california	`masc.+= 1`
Cat: un tauró fereix un **nen** de 13 anys mentre buscava llagostes a califòrnia	`masc.+= 1`
Sentence 2: Multiple translations and variation in part of speech	
Eng: [...] something increasingly demanded by younger shoppers.	`unspecified.+= 0`
Cat: [...] un aspecte cada cop més demanat pels consumidors més **joves**.	`unspecified.+= 1`
Sentence 3: Robust to typographic differences	
Eng: **mother**-of-three willoughby and **husband** dan baldwin have been close to jones and his **wife**	`fem.+= 2,masc.+= 1`
Cmn: [...]个孩子的母亲的威洛比及其丈夫 dan baldwin 十年来与琼斯及其妻子 tara保持[...]	`fem.+= 2,masc.+= 1`
Sentence 4: Synonyms	
Eng: [...] the owner of the lloyds pharmacy chain, for £125m, three years ago.	`masc.+= 0`
Vie: [...] chù sõ hũu cũa chuõi nhà thuòc lloyds, vói giá 125 triẽu bàng vào **ba** năm trùózc.	`masc.+= 1`

Table 2: Selected examples of gender representation across parallel sentences between English and multiple target languages (based on the NTREX dataset). Detected gendered nouns in bold/underlined. We indicate the counter incremented by the pipeline for the three gender classes (feminine, masculine and unspecified) next to each sentence when there is at least a match in one of the languages.

Common Crawl includes a large diversity of domains from the Web.

Comparing Genders across Languages In addition, we find a large variability across languages. Some languages like Belarus (`bel`) and Swedish (`swe`) are highly skewed toward the Masculine gender class, while other languages are much more balanced such as Mandarin Chinese (`cmn`) or Hindi (`hin`).

We note that gender distribution cannot be compared across languages quantitatively. Indeed, first, our lexicon is based by design on nouns that are not entirely parallel across languages. Second, our metric highly depends on the number of words in each dataset, which is not comparable across all languages due to their differences in morphology and syntax. However, as discussed below (§ 5.2), our pipeline allows us to highlight qualitative differences in how gender is encoded in different languages.

5.2 Qualitative Analysis: Gender representation variation in parallel data

To understand the cause of these gender representation differences across languages, we present several examples in Table 2. We dicuss them here:

- Omission of words: When comparing English with Romance languages, we observe cases where the gendered word is omitted in English while being translated as a masculine noun in the target language, like Spanish or Catalan. This leads to larger gender representation gaps in these languages.
- Multiple translations and part-of-speech: Sentence 2 shows the impact of how a single English word corresponds to multiple words in other languages. The unspecified word "kid" is translated in 10 words in Catalan: unspecified "jove, criatura"; feminine "minyona, menuda, nena, marreca"; masculine, "minyó, menut, nen, marrec", augmenting the coverage in that second language. In addition, some words in Catalan have multiple part-of-speech, like "jove, menuda, menut" which can act as nouns or adjectives.
- Sentence 3 illustrates that even with typologically different languages such as English and Mandarin Chinese, our lexical matching approach successfully highlights cases where gender is preserved across languages.
- Finally, in Sentence 4, we illustrate the limit of the context-free approach. Indeed, the noun “ba” means both *father* and *three* in Vietnamese, leading to over-estimating the masculine class on some samples.

In summary, the differences in gender representation across languages point to four distinct phe-

nomena: First, the inherent limit of our context-free lexical approach. Gender is, in some cases, incorrectly estimated by a by-design restricted lexical-matching method (e.g., Sentence 4). Second, different domain distributions may lead to diverse gender representation. As reported in the previous section, for some languages, the gender scores highly vary depending on the domains (e.g., News vs. Web crawled data). This suggests that when we analyze non-parallel data, the domain may be a prevalent factor that explains gender representation differences across languages. Third, as we observe when analyzing parallel data, gender representation differences may come from biases in the translation itself. For instance, in Sentence 1, the translation explicitly encoded the masculine gender in Spanish and Catalan while being gender unspecified in English. Other translations could have preserved the gender. Fourth, the way gender is encoded is, partly at least, unique to each language. Some languages are inherently biased toward the masculine gender (e.g. "padres", which may mean both *fathers* and *parents* in Spanish). Other languages do not always have genderless nouns. For instance, *siblings* can only be translated onto Lithuanian as "broliai ir seserys" *Brothers and Sisters*.

6 Conclusion

In this work, we presented GENDER-GAP, a large scale multilingual pipeline to compute gender distribution across 55 languages. We find that broadly used datasets are biased toward masculine gender. Based on this finding, our primary recommendation for multilingual NLP practitioner is to report the gender distribution along with the performance score. This allows reader and systems adopters to be aware of these biases in order to integrate this in their system deployment. Secondly, based on our multilingual lexicon, many directions could be taken to mitigate biases in the performance of the systems (due to biases in the data). Qian et al. (2022) developed a perturbation-based technique to build NLP systems that are less biased toward specific group. We envision using our multilingual lexicon to adapt this technique beyond English.

Limitations

English-centric We designed the list of gendered nouns starting from the English language and then scaled it to multiple languages. This means that our approach may cover incompletely the nuances in different language families regarding gender or only cover them partially and from an English-centric perspective.

Non-Binary Gender Modeling To favor scalability across 55 languages, we chose to use a three gender class lexicon. However, this restrict our approach to binary genders (masculine and feminine) and we only measure imperfectly non-binary genders distribution (Haynes et al., 2001; Herdt, 2020) with the "unspecified" class. We leave for future work the refinement of our lexical categories in order to measure more granularly genders across languages.

Lexical Matching The core assumption of this work is that our predefined lexicon defined in Section 3.2 gives us a proxy to account for gender distributions in large datasets. Although our lexicon is obviously not exhaustive, it is simple enough to scale to highly multilingual environments. Future work could consider other types of nouns (beyond family relations or persons) such as gendered occupations nouns, pronouns, etc.

References

Lauren Ackerman 2019. Syntactic and cognitive issues in investigating gendered coreference. *Glossa: a journal of general linguistics*, 4.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

M.G. Bulmer. 1979. *Principles of Statistics*. Dover Books on Mathematics Series. Dover Publications.

Judith Butler. 2011. *Bodies that matter: On the discursive limits of sex*. Taylor & Francis.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Prafulla Kumar Choubey, Anna Currey, Prashant Mathur, and Georgiana Dinu. 2021. GFST: Gender-filtered self-training for more accurate gender in translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1640–1654, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Greville G. Corbett. 1991. *Gender*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Greville G. Corbett. 2013. Number of genders (v2020.3). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Ander Corral and Xabier Saralegi. 2022. Gender bias mitigation for NMT involving genderless languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 165–176, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Marta Costa-jussà, Christine Basta, Oriol Domingo, and André Rubungo. 2022. Occgen: Selection of real-world multilingual parallel data balanced in gender within occupations. In *Advances in Neural Information Processing Systems*, volume 35, pages 1445–1457. Curran Associates, Inc.

Marta R Costa-jussà, Pierre Andrews, Eric Smith, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Daniel Licht, and Carleigh Wood. 2023. Multilingual holistic bias: Extending descriptors and patterns to unveil demographic biases in languages at scale. *arXiv preprint arXiv:2305.13198*.

Marta R. Costa-jussà and Adrià de Jorge. 2020. Fine-tuning neural machine translation on gender-balanced datasets. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 26–34, Barcelona, Spain (Online). Association for Computational Linguistics.

Marta Ruiz Costa-jussà, Pau Li Lin, and Cristina España-Bonet. 2019. Gebiotoolkit: Automatic extraction of gender-balanced multilingual corpus of wikipedia biographies. In *International Conference on Language Resources and Evaluation*.

Council of Europe. 2023. Sex and gender. `https://www.coe.int/en/web/gender-matters/sex-and-gender`. [Accessed: July 17, 2023].

Brian Duignan. 2023. gender continuum. Encyclopedia Britannica.

Joel Escudé Font and Marta R. Costa-jussà. 2019. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154, Florence, Italy. Association for Computational Linguistics.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

J. R. Foulds, R. Islam, K. Keya, and S. Pan. 2020. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1918–1921, Los Alamitos, CA, USA. IEEE Computer Society.

Felicity Haynes, Tarquam McKenna, and E McWilliam. 2001. *Unseen genders: Beyond the binaries*. Peter Lang Publishing.

M. Hellinger and H. Bussmann. 2001. *Gender Across Languages: The Linguistic Representation of Women and Men*. Number vol. 2 in Gender Across Languages: The Linguistic Representation of Women and Men. J. Benjamins.

Gilbert Herdt. 2020. *Third sex, third gender: Beyond sexual dimorphism in culture and history*. Princeton University Press.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "you sound just like your father" commercial machine translation systems include stylistic biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690, Online. Association for Computational Linguistics.

Ruth Kramer. 2020. Grammatical gender: A close look at gender assignment across languages. *Annual Review of Linguistics*, 6:45–66.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6).

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. 2020. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32:6363–6381.

Jennifer L. Prewitt-Freilino, T Andrew Caswell, and Emmi K. Laakso. 2012. The gendering of language: A comparison of gender equality in countries with gendered, natural gender, and genderless languages. *Sex Roles*, 66:268–281.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Michael Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer NLP. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9496–9521, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Adithya Renduchintala, Denise Diaz, Kenneth Heafield, Xian Li, and Mona Diab. 2021. Gender bias amplification during speed-quality optimization in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 99–109, Online. Association for Computational Linguistics.

Adithya Renduchintala and Adina Williams. 2022. Investigating failures of automatic translationin the case of unambiguous gender. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3454–3469, Dublin, Ireland. Association for Computational Linguistics.

Danielle Saunders, Rosie Sallis, and Bill Byrne. 2020. Neural machine translation doesn't translate gender coreference right unless you make it. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 35–43, Barcelona, Spain (Online). Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "I'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Artūrs Stafanovičs, Toms Bergmanis, and Mārcis Pinnis. 2020. Mitigating gender bias in machine translation with target gender annotations. In *Proceedings of the Fifth Conference on Machine Translation*, pages 629–638, Online. Association for Computational Linguistics.

Karolina Stanczak and Isabelle Augenstein. 2021. A survey on gender bias in natural language processing. *arXiv preprint arXiv:2112.14168*.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Angelina Wang and Olga Russakovsky. 2021. Directional bias amplification. In *International Conference on Machine Learning*, pages 10882–10893. PMLR.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Lang	Feminine	Masculine	Unspecified	Δ(| Fem.-Masc. |) (ste.)	# words	% matched sentences
			Flores DevTest.			
eng	0.121	0.065	**0.379**	0.056 (0.0003)	23211	11.2
arb	0.051	0.047	**0.094**	0.004 (0.0002)	25549	4.1
asm	0.056	**0.102**	0.093	0.046 (0.0003)	21610	4.5
bel	0.161	0.434	**0.444**	0.274 (0.0005)	21174	12.7
ben	0.076	**0.204**	0.142	0.128 (0.0004)	21101	7.2
bul	0.083	**0.258**	0.114	0.175 (0.0004)	22834	9.1
cat	0.115	**0.154**	0.146	0.038 (0.0003)	26005	9.4
ces	0.113	**0.385**	0.153	0.271 (0.0005)	20284	10.6
ckb	0.052	0.119	**0.152**	0.066 (0.0003)	21073	4.3
cmn	0.101	0.042	**0.794**	0.059 (0.0002)	23676	17.6
cym	0.104	0.046	**0.146**	0.058 (0.0002)	26013	6.4
dan	0.129	0.045	**0.160**	0.085 (0.0003)	22471	6.3
deu	0.114	0.059	**0.301**	0.055 (0.0003)	21922	9.2
ell	0.118	**0.261**	0.253	0.143 (0.0004)	24548	12.8
est	0.116	0.099	**0.519**	0.017 (0.0003)	18107	11.0
fin	0.116	0.086	**0.147**	0.031 (0.0004)	16314	4.9
fra	0.082	0.089	**0.234**	0.007 (0.0003)	26910	9.6
gle	0.038	0.053	**0.479**	0.015 (0.0002)	26517	12.3
hin	0.048	0.032	**0.104**	0.016 (0.0002)	25094	3.8
hun	0.040	**0.250**	0.060	0.210 (0.0004)	19977	6.0
ind	0.179	**0.468**	0.193	0.289 (0.0006)	20728	14.5
ita	0.082	0.168	**0.223**	0.086 (0.0003)	25583	10.2
jpn	0.113	0.061	**0.716**	0.052 (0.0002)	31000	20.4
kan	0.086	0.032	**0.102**	0.054 (0.0002)	18593	3.1
kat	**0.097**	0.029	0.068	0.068 (0.0002)	20527	3.0
khk	0.274	**0.874**	0.270	0.599 (0.0007)	21861	22.6
kir	0.134	0.194	**0.482**	0.060 (0.0004)	20120	12.7
kor	0.037	**0.055**	0.012	0.018 (0.0002)	16341	1.7
lit	**0.140**	0.088	0.125	0.052 (0.0003)	19246	5.4
lug	0.084	0.023	**0.606**	0.061 (0.0002)	21457	12.6
mar	**0.060**	0.044	0.055	0.016 (0.0002)	18281	2.5
mlt	**0.661**	0.179	0.191	0.482 (0.0005)	25104	18.3
nld	0.113	0.071	**0.236**	0.042 (0.0003)	21229	7.5
pan	0.105	**0.127**	0.087	0.022 (0.0003)	27651	6.5
pes	0.166	0.116	**0.310**	0.050 (0.0003)	24157	10.0
pol	0.137	0.061	**0.544**	0.076 (0.0003)	21143	13.4
por	0.103	0.078	**0.338**	0.025 (0.0003)	24269	10.9
ron	0.100	0.092	**0.240**	0.008 (0.0003)	25046	8.9
rus	0.117	**0.117**	0.098	0.000 (0.0003)	21431	5.4
slk	0.113	0.054	**0.508**	0.059 (0.0003)	20292	11.5
slv	0.069	0.032	**0.069**	0.037 (0.0002)	21586	3.3
spa	0.104	0.201	**0.260**	0.097 (0.0003)	26896	12.3
swe	0.119	0.176	**0.200**	0.057 (0.0004)	20969	8.9
swh	0.225	0.213	**0.689**	0.013 (0.0004)	23964	20.4
tam	**0.168**	0.101	0.123	0.067 (0.0003)	17862	4.5
tel	0.092	**0.140**	0.122	0.049 (0.0004)	16373	3.9
tgl	0.075	0.041	**0.373**	0.034 (0.0002)	29518	11.1
tha	0.156	0.038	**0.439**	0.118 (0.0003)	28922	12.7
tur	0.287	0.270	**0.293**	0.017 (0.0005)	17775	8.4
urd	0.074	**0.320**	0.234	0.245 (0.0004)	26887	9.2
uzn	0.156	0.076	**0.260**	0.080 (0.0003)	21181	8.3
vie	0.139	0.301	**1.441**	0.162 (0.0004)	25263	30.6
yue	0.093	0.040	**0.837**	0.053 (0.0002)	24728	19.1
zul	0.394	0.059	**0.653**	0.335 (0.0005)	18532	17.0
avg.	0.128	0.144	**0.302**	0.097 (0.0003)	22572	10.1

Table 3: % Gender Distribution in FLORES-200 dataset (NLLB Team et al., 2022). We **bold** the most represented gender class, and underline the second most represented gender class for each language. We report Δ the gender gap defined as the absolute difference between the Feminine and Masculine scores along with the standard error (ste.). % matched sentences refers to the coverage of our pipeline (cf. § 5.1).

Lang	Feminine	Masculine	Unspecified	Δ(| Fem.-Masc. |) (ste.)	# words	% matched sentences
			NTREX			
eng	0.166	0.203	**0.379**	0.037 (0.0003)	48254	15.5
arb	0.105	0.107	**0.206**	0.002 (0.0002)	51388	8.7
bel	0.224	**0.574**	0.397	0.350 (0.0004)	44597	16.9
ben	0.131	0.212	**0.311**	0.081 (0.0003)	40505	11.3
bul	0.122	**0.270**	0.095	0.148 (0.0003)	49283	10.5
cat	0.195	**0.272**	0.235	0.077 (0.0003)	54401	15.6
ces	0.248	**0.454**	0.190	0.206 (0.0004)	43623	16.3
ckb	0.054	0.167	**0.244**	0.113 (0.0002)	42554	6.5
cmn	0.193	0.149	**0.944**	0.044 (0.0003)	50326	24.8
cym	0.086	**0.164**	0.154	0.078 (0.0002)	52540	8.8
dan	0.184	0.177	**0.186**	0.007 (0.0003)	45684	10.7
deu	0.162	0.192	**0.276**	0.030 (0.0003)	46398	12.3
ell	0.141	**0.344**	0.170	0.203 (0.0003)	51204	14.4
est	0.212	0.328	**0.458**	0.116 (0.0004)	37794	15.8
fin	0.158	0.181	**0.196**	0.024 (0.0003)	33617	7.9
fra	0.140	0.208	**0.258**	0.068 (0.0003)	54336	13.9
gle	0.081	0.135	**0.493**	0.054 (0.0002)	54205	16.2
hin	0.103	0.092	**0.147**	0.011 (0.0002)	55207	8.1
hun	0.110	**0.140**	0.072	0.030 (0.0002)	42834	6.6
ind	0.195	**0.581**	0.213	0.386 (0.0004)	45071	18.1
ita	0.166	**0.301**	0.229	0.135 (0.0003)	51884	14.8
jpn	0.209	0.201	**0.868**	0.008 (0.0002)	59704	25.2
kan	0.115	0.101	**0.131**	0.014 (0.0002)	36574	4.9
kat	**0.198**	0.140	0.103	0.058 (0.0002)	39912	5.1
kir	0.209	0.181	**0.310**	0.028 (0.0003)	38682	12.0
kor	0.040	**0.062**	0.059	0.022 (0.0002)	32204	2.5
lit	0.187	**0.216**	0.153	0.029 (0.0003)	41190	9.4
mar	**0.089**	0.056	0.069	0.033 (0.0002)	35980	3.6
mlt	**0.795**	0.212	0.284	0.583 (0.0004)	51466	24.7
nld	0.190	0.194	**0.196**	0.004 (0.0003)	48003	11.2
pan	0.150	**0.176**	0.100	0.026 (0.0002)	53845	9.9
pol	0.242	0.211	**0.525**	0.030 (0.0003)	42638	17.9
por	0.160	0.228	**0.244**	0.067 (0.0003)	50482	13.8
ron	0.152	0.191	**0.367**	0.039 (0.0002)	54463	15.5
rus	0.171	**0.210**	0.089	0.039 (0.0003)	46295	8.5
slk	0.248	0.216	**0.420**	0.033 (0.0003)	43063	16.0
slv	**0.093**	0.084	0.077	0.009 (0.0002)	45339	4.8
spa	0.162	0.297	**0.344**	0.135 (0.0003)	52579	15.9
swe	0.156	**0.265**	0.240	0.109 (0.0003)	42980	12.3
tam	**0.308**	0.273	0.068	0.035 (0.0002)	36960	7.0
tel	0.118	**0.213**	0.086	0.095 (0.0003)	31427	5.0
tha	0.418	0.128	**0.870**	0.290 (0.0003)	57923	23.1
tur	0.227	0.183	**0.252**	0.044 (0.0003)	36163	8.1
vie	0.146	0.633	**2.166**	0.487 (0.0004)	52577	45.7
yue	0.133	0.173	**0.933**	0.041 (0.0002)	54233	26.6
avg.	0.180	0.224	0.329	0.099 (0.0003)	46231	13.4

Table 4: % Gender Distribution in NTREX data (Federmann et al., 2022). We **bold** the most represented gender class, and underline the second most represented gender class for each language. We report Δ the gender gap defined as the absolute difference between the Feminine and Masculine scores along with the standard error (ste.). % matched sentences refers to the coverage of our pipeline (cf. § 5.1).

Lang	Feminine	Masculine	Unspecified	Δ(| Fem.-Masc. |) (ste.)	# words	% matched documents
			CommonCrawl			
eng	0.120	0.115	**0.243**	0.005 (0.0000)	2529756	9.4
arb	0.101	**0.106**	0.085	0.005 (0.0000)	6078083	9.5
bel	0.122	**0.447**	0.358	0.325 (0.0000)	2430561	14.1
ben	0.158	**0.199**	0.140	0.041 (0.0000)	4603054	14.4
bul	0.072	**0.145**	0.142	0.073 (0.0000)	2708232	7.7
cat	0.079	0.141	**0.152**	0.062 (0.0000)	3157729	9.1
ces	0.117	0.146	**0.165**	0.030 (0.0000)	2366804	7.9
ckb	0.108	0.049	**0.124**	0.059 (0.0000)	5341945	10.2
cmn	0.170	0.097	**0.519**	0.072 (0.0000)	5484451	23.8
cym	0.079	0.082	**0.164**	0.003 (0.0000)	2777579	7.4
dan	0.182	0.102	**0.201**	0.080 (0.0000)	2310993	7.9
deu	0.144	0.099	**0.187**	0.044 (0.0000)	2148705	6.8
ell	0.068	**0.143**	0.142	0.075 (0.0000)	2855903	7.7
est	0.112	0.152	**0.429**	0.040 (0.0000)	1943773	10.3
fin	**0.294**	0.201	0.155	0.094 (0.0001)	1621020	7.3
fra	0.110	0.136	**0.151**	0.025 (0.0000)	2857434	8.3
gle	0.044	0.101	**0.406**	0.057 (0.0000)	2634719	12.2
hin	**0.176**	0.124	0.065	0.052 (0.0000)	2675603	7.4
hun	0.058	**0.097**	0.075	0.038 (0.0000)	2572506	4.5
ind	0.183	**0.367**	0.184	0.185 (0.0000)	2227691	12.1
ita	0.131	**0.195**	0.070	0.064 (0.0000)	2961219	8.2
jpn	0.858	0.724	**0.963**	0.134 (0.0000)	5964414	27.4
kan	**0.103**	0.094	0.093	0.009 (0.0000)	3772755	7.0
kat	**0.129**	0.089	0.116	0.040 (0.0000)	3977699	6.2
khk	0.301	**0.948**	0.248	0.647 (0.0000)	4996882	32.1
kir	0.269	**0.308**	0.270	0.039 (0.0000)	3895597	20.0
kor	0.032	0.047	**0.047**	0.015 (0.0000)	2364450	2.4
lit	0.148	0.117	**0.243**	0.031 (0.0000)	2293338	8.4
mar	**0.133**	0.112	0.051	0.021 (0.0000)	1531197	3.8
mlt	**0.554**	0.179	0.213	0.375 (0.0001)	2437212	20.0
nld	0.127	0.101	**0.201**	0.027 (0.0000)	1921934	6.3
pan	0.236	**0.308**	0.074	0.072 (0.0000)	6772503	22.4
pes	1.459	1.425	**1.514**	0.034 (0.0000)	3881584	14.7
pol	0.175	0.074	**0.290**	0.101 (0.0000)	2453053	9.9
por	0.110	**0.160**	0.158	0.050 (0.0000)	2846706	9.2
ron	0.207	0.138	**0.257**	0.068 (0.0000)	2555624	10.1
rus	0.107	**0.139**	0.117	0.031 (0.0000)	2565203	6.4
slk	0.111	0.066	**0.324**	0.045 (0.0000)	2269033	8.7
slv	0.057	0.071	**0.142**	0.014 (0.0000)	2373967	5.3
spa	0.122	**0.255**	0.183	0.133 (0.0000)	3046193	11.7
swe	0.179	**0.372**	0.157	0.193 (0.0000)	2346273	11.5
swh	0.221	0.194	**0.492**	0.027 (0.0000)	2385794	19.9
tam	**0.766**	0.676	0.073	0.091 (0.0000)	1691612	11.7
tel	0.127	**0.165**	0.056	0.038 (0.0000)	1277513	3.5
tgl	0.145	0.108	**0.419**	0.038 (0.0000)	5035687	21.2
tha	0.735	0.107	**0.932**	0.628 (0.0000)	7142646	28.9
tur	**0.228**	0.202	0.215	0.027 (0.0000)	2293026	7.3
uzn	0.119	0.077	**0.280**	0.042 (0.0000)	2973725	9.5
avg.	0.212	0.260	0.251	0.136 (0.0003)	3088848	12.0

Table 5: % Gender Distribution in a Common Crawl sample. We **bold** the most represented gender class, and underline the second most represented gender class. We report Δ the gender gap defined as the absolute difference between the Feminine and Masculine scores along with the standard error (ste.). % matched documents refers to the coverage of our pipeline (cf. § 5.1).

Language Code	Language
arb_Arab	Modern Standard Arabic
asm_Beng	Assamese
bel_Cyrl	Belarusian
ben_Beng	Bengali
bul_Cyrl	Bulgarian
cat_Latn	Catalan
ces_Latn	Czech
ckb_Arab	Central Kurdish
cmn_Hans	Mandarin Chinese (simplified script)
cym_Latn	Welsh
dan_Latn	Danish
deu_Latn	German
ell_Grek	Greek
eng_Latn	English
est_Latn	Estonian
fin_Latn	Finnish
fra_Latn	French
gle_Latn	Irish
hin_Deva	Hindi
hun_Latn	Hungarian
ind_Latn	Indonesian
ita_Latn	Italian
jpn_Jpan	Japanese
kat_Geor	Georgian
khk_Cyrl	Halh Mongolian
kir_Cyrl	Kyrgyz
lit_Latn	Lithuanian
lug_Latn	Ganda
lvs_Latn	Standard Latvian
mar_Deva	Marathi
mlt_Latn	Maltese
nld_Latn	Dutch
pan_Guru	Eastern Panjabi
pes_Arab	Western Persian
pol_Latn	Polish
por_Latn	Portuguese
ron_Latn	Romanian
rus_Cyrl	Russian
slk_Latn	Slovak
slv_Latn	Slovenian
spa_Latn	Spanish
swe_Latn	Swedish
swh_Latn	Swahili
tam_Taml	Tamil
tha_Thai	Thai
tur_Latn	Turkish
ukr_Cyrl	Ukrainian
urd_Arab	Urdu
uzn_Latn	Northern Uzbek
vie_Latn	Vietnamese
yue_Hant	Yue Chinese (traditional script)
kan_Knda	Kannada
tel_Telu	Telugu
tgl_Latn	Tagalog
zul_Latn	Zulu

Table 6: The 55 languages analyzed in this work, subselected from the 200 NLLB languages (NLLB Team et al., 2022).

Towards better evaluation for Formality-Controlled English-Japanese Machine Translation

Edison Marrese-Taylor[1,2], Pin-Chen Wang[2], Yutaka Matsuo[2]
[1] National Institute of Advanced Industrial Science and Technology
[2] Graduate School of Engineering, The University of Tokyo
{emarrese,wangpinchen,matsuo}@weblab.t.u-tokyo.ac.jp

Abstract

In this paper we propose a novel approach to automatically classify the level of formality in Japanese text, using three categories (formal, polite, and informal). We introduce a new dataset that combine manually-annotated sentences from existing resources, and formal sentences scrapped from the website of the House of Representatives and the House of Councilors of Japan. Based on our data, we propose a Transformer-based classification model for Japanese, which obtains state-of-the-art results in benchmark datasets. We further propose to utilize our classifier to study the effectiveness of prompting techniques for controlling the formality level of machine translation (MT) using Large Language Models (LLM). Our experimental setting includes a large selection of such models and is based on an En→Ja parallel corpus specifically designed to test formality control in MT. Our results validate the robustness and effectiveness of our proposed approach and while also providing empirical evidence suggesting that prompting LLMs is a viable approach to control the formality level of En→Ja MT using LLMs.

1 Introduction

Communication by way of natural language often includes indicators for respect to acknowledge the hierarchy, interpersonal relationship, and power dynamics of the participants in a conversation or written text. In this context, formality or honorifics refers to the set of linguistic features used to establish the degree of respect and deference conveyed in a given context.

Naturally, these phenomena exhibit significant variation across different languages and cultures (Biber and Conrad, 2019). While many European languages emphasize formality through the use of standard grammar, more complicated sentence structures (active, passive, use of clauses, etc.), or more advanced and complex choice of vocabulary and phrases, the Japanese language has its own formality system. This system, named Keigo (敬語), requires users to identify the status or the relationship with the interlocutor, is strict, following a standard grammar format (Fukada and Asato, 2004), and can generally be divided into four different categories (Aoki et al., 2007), as follows.

- **Regular (jyotai,** 常体**)**: a form that is often used in, but not limited to a daily conversation with only people one is familiar with or people who are in the equivalent social status.
- **Polite (teineigo,** 丁寧語**)**: a form that is generally used throughout the whole Japanese society to create some distance between one another. Although this form does not indicate the amount of respect one holds toward others, it helps deliver messages in a polite way that will not be offensive on any occasion.
- **Respectful (sonkeigo,** 尊敬語**)**: a form that shows extensive respect, which is used to maximize the preeminence of the interlocutor.
- **Humble (kenjyougo,** 謙譲語**)**: a form that specifies humbleness, which is used by the Japanese speakers to minimize their own value in order to highlight the greatness of the interlocutor.

In this context, what makes Japanese formality stand out is that it allows to convert any sentences from one style to another by simply adjusting the tense of the verb (Aoki et al., 2007), while maintaining the original meaning, word choice, and sentence structure.

Additionally, the system follows one additional rule (Aoki et al., 2007), where one can always mix the four forms together in one paragraph. The more respectful form one uses in a sentence or a paragraph, the more courtesy one states toward one's interlocutor. Similarly, the more humble form

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 551–560
December 6–7, 2023.

one uses, the more modest one is in the conversation. However, it is also emphasized that when containing too many formal terms in a sentence, the sentence will become annoying and considered inappropriate in Japanese social rules (Aoki et al., 2007).

Given the importance of formality in language generation systems such as machine translation (MT), the ability to control formality and honorifics is a critical factor in achieving accurate and appropriate results. In particular, for the Japanese language, failure in recognizing and incorporating levels of formality can result in unnatural, impolite, or disrespectful translations, which can impede effective communication across diverse linguistic and cultural contexts (Fukada and Asato, 2004). We therefore think that developing and refining MT models that can accurately control honorific levels is crucial for this language. Although formality-controlled machine translation (FCMT) has gained popularity for languages like English (Niu and Carpuat, 2020), there is a substantial lack of resources to tackle the formality problem for Japanese, which extends to the more fundamental task of formality detection.

In light of this issue, we focus on developing resources to improve formality detection in Japanese. We begin by uncovering several flaws on existing corpora for the task, including issues such as the presence of ungrammatical sentences, as well as wrong formality labels. To alleviate these issues, we introduce new resources for Japanese formality detection which consists of manually-labeled sentences annotated with three formality classes (informal, polite, and formal). We propose this three-way setting in opposition to existing resources which are annotated using binary labels, to better approximate the nature of formality of the Japanese language. As existing resources (Nadejde et al., 2022; Liu and Kobayashi, 2022) lacked data for the formal label, a part of our dataset is constructed with sentences sampled from these sources and with text obtained from meeting minutes from committees of the House of Representatives and the House of Councilors of Japan [1].

Furthermore, as language generation models based on Large Language Models (LLMs) have recently been able to attain substantial performance improvements on language generation benchmarks, we note that the lack of a consistent evaluation method makes it difficult to verify to what extent such models can perform formality control. In MT, current studies mainly rely on human assessment or simple models. For example, Feely et al. (2019) and Nadejde et al. (2022) use rule-based methods where lists of grammatical rules are combined with pattern-matching to perform classification. Though formality-level classifiers for Japanese based on machine learning have been proposed in the past (Rippeth et al., 2022; Liu and Kobayashi, 2022), so far this has been without focus on MT or lacked proper evaluation.

In consideration of the above issue, in this paper we propose a novel approach, based on machine learning, to evaluate the ability of En→Ja MT models to perform formality-control. Concretely, we use our dataset to train a robust Transformer-based classifier that leverages a masked-language model, which is able to obtain state-of-the-art performance on our dataset and on existing Japanese formality detection benchmarks. Following recent work relying on machine learning models to evaluate language generation, such as BERTScore (Zhang et al., 2019), and MT models, such as COMET (Rei et al., 2020a), we present an empirical study using our classifier to evaluate the zero-shot ability of several state-of-the-art LLMs to perform formality control.

Our results validate the effectiveness of our proposed approach and show that, compared to existing evaluation techniques that rely on rules and expression-matching, it offers a robust, reliable, and accurate evaluation metric for formality-controlled MT systems. We further demonstrate the ability of LLMs to generate sequences with varying levels of formality through the use of well-designed prompts, concretely showing that both GPT-3 and ChatGPT can attain a formality control accuracy of approximately 90%, and ultimately suggesting that prompting LLMs can result in better formality control performance than fine-tuned MT models. We release or data and trained models[2] to encourage further research on this topic.

2 Related Work

To the best of our knowledge, previous work on formality detection for Japanese is relatively recent and limited in scope, with only two existing resources. On the one hand, we find the Japanese portion of the CoCoA-MT (Nadejde et al., 2022)

[1] https://kokkai.ndl.go.jp/

[2] https://github.com/epochx/japanese-formality

dataset, which was released for the 2022 Shared Task on Formality Control at IWSLT (Anastasopoulos et al., 2022) and contains a total of 1,600 parallel English-Japanese sentences (1,000 for training, and 600 for testing). The source data for this corpus comes from Topical-Chat4 (Gopalakrishnan et al., 2019), as well as Telephony and Call Center data, containing text-based conversations about various topics. For each segment, one reference translation for each formality level (formal and informal) were collected. For the Japanese translations, informal was mapped to jyoutai, and formal was mapped to teineigo, sonkeigo and/or kenjyougo.

On the other hand, we find the recently-released KeiCO corpus (Liu and Kobayashi, 2022), which has a total of 10,007 examples across the four forms of the Japanese formality system (Levels 1 to 4, according to the paper). It additionally contains detailed information about the presence of level-related honorifics —a sentence may contain markers for multiple levels of politeness— the social relationship between the speaker and the listener, and conversational situations or topics. To obtain this data, 40 native Japanese volunteers were asked to regenerate a total of 3,000 sentences coming from machine translation, dialogue systems, and semantic analysis systems, by filling in blanks with honorifics.

The two datasets mentioned above have been used to train Transformer-based classifiers. Liu and Kobayashi (2022) rely on Japanese-BERT (Suzuki and Takahashi, 2021), while the submission of Rippeth et al. (2022) for the 2022 Shared Task on Formality Control at IWSLT relied on XLM-R (Conneau et al., 2020).

Our work is also related to FCMT. In this context, recent approaches have relied on formality-annotated parallel corpora such as CoCoA-MT, early work on this task resorted to other resources such as rule-based generation of synthetic data for English-Japanese (Feely et al., 2019) and English-German (Sennrich et al., 2016), as well as synthetic supervision by means of multi-tasking (formality classification and machine translation). We also find that these studies rely on rule-based simple approaches to measure the accuracy of formality control in the translation, or directly perform human assessment. For example, the FSMT approach English-French by Niu et al. (2017) conducted a human study in which they assigned translation pairs for human annotators. Neural CFMT models for English-Japanese (Feely et al., 2019) and English-German (Sennrich et al., 2016) depend on rule-based classifiers, where grammatical rules for the language are listed and matched.

The recent rise of LLMs has enabled models to perform certain language generation tasks in zero-shot or few-shot manner (Brown et al., 2020), or by means of prompts. Some of these capabilities have been further enhanced by means of prompt-based training (Sanh et al., 2022), where zero-shot generalization is induced by explicit multitask learning. This work is relevant to our paper, as we test the ability of several such models to perform zero-shot FCMT. Our study also considers multilingual efforts in Neural MT, admittedly also a kind of LLM, where we look at M2M100 (Fan et al., 2021) and NLLB200 (Costa-jussà et al., 2022)

Finally, we also find recent work on using few-shot prompting-based techniques to control the formality level of English-German Machine Translation (Garcia et al., 2023). Also, Pu and Demberg (2023) recently performed an in-depth study of the capabilities of ChatGPT to generate text in different styles, including formal/informal labels, showing that the model sometimes incorporates factual errors or hallucinations when adapting the text to suit a specific style.

3 A robust classifier for Japanese Formality

3.1 Data

The size and quality of the datasets are vital requisites to maximize the performance of the machine learning models (Mohri et al., 2018). As one goal of our work is to train a robust classifier for Japanese formality, we look at two main issues. In contrast to existing resources, which either offer limited flexibility by simplifying the dynamics of Japanese formality into two classes (Nadejde et al., 2022), or are too specific by exactly following the grammar (Liu and Kobayashi, 2022), we propose a compromise between these and divide the Japanese language into three categories based on the four formality levels and their corresponding applied situations: (1) “Informal” (for regular tense), (2) “Polite” (for polite tense or teineigo), and (3) Formal (for respectful and humble tenses). Below, we detail how we transformed existing datasets for our purposes, created new resources when necessary, and how we constructed a final curated corpus to train our model.

RECOCOA-MT As we divided Japanese formality into 3 classes, this suggested that the re-utilization of the Japanese portion CoCoA-MT corpus required a transformation of the labels, so we began by analyzing the data. During this stage, we found that many of the examples of the parallel corpus contain broken sentences, while in many cases other sentences do not have an understandable Japanese meaning. Based on these observations, we decided to re-annotate the data and recruited volunteer Japanese native speakers to proceed[3]. During the re-annotation procedure, we confirmed that 44 out of the 1,000 training examples were mislabeled. After re-annotation and filtering, 520 sentences are labeled as informal, 464 sentences as polite, and 12 sentences as formal.

KOKAI As seen above, the re-annotation of COCOA-MT showed that the label distribution in this dataset was heavily skewed away from the formal label, which suggested that more data for this particular level of formality was required. Noting that Japanese political committees tend to rely on language that is considered formal, or at least polite, with very little informal syntax, we proceeded to collect all the meeting minutes from the Japanese Congress (House of Representatives of Japan and the House of Councilors of Japan) from 1947 to 2022. In total, we obtained 64,630 sentences with 23,672 paragraphs, excluding 11,805 broken sentences which are mostly the names, dates, or titles of the committees or the list of participants. We surmise some of these broken sentences, as well as the informal sentences that we observed upon close examination, are likely interrupted utterances that occurred during the sessions. Despite the overall formal nature of the source of data, to ensure the quality of the labels we use for training, we randomly selected 1,360 examples from the raw data and ask our volunteer Japanese native speakersfootnote:annotators to annotate the examples following the same procedure as before. As a result, we obtain 137 informal examples, 760 polite examples, and 463 formal examples.

DAILY We collected 200 sentences sampled from Japanese news, novels, textbooks, business letters and academic documents. The dataset is well-balanced across our three labels with 65, 67, and 68 sentences for the informal, polite, and formal classes, respectively. We use this small corpus mainly for preliminary experiments, but also include these examples in the data used to train our model, as explained below.

KEICO We note that according to Liu and Kobayashi (2022), both respectful and humble tenses are used for the proposed formality Levels 1 and 2. We therefore simply map these two classes to our formal class, to make the annotations compatible with our setting

Using these sources of data except the KEICO, we prepared a first training set consisting of 1,000 examples (426 informal, 501 polite sentences, and 273 formal), leaving a total of additional 200 examples left for development purposes. Though KOKAI has been collected to specifically cover for the lack of annotated data for the formal label in RECOCOA-MT, because the contents are highly related to politics and other related domains, we hypothesize that by only utilizing examples derived for this dataset for training may lead to models that may fail to generalize well to other situation where the respectful and humble tenses are also utilized. To that end, for the initial training set we purposely omit examples from KEICO, which contains formal examples from a diverse set of domains, and build a second training set that relies on examples taken from this corpus to balance the topic distribution. We take a total of 2K examples from KEICO (with 530, 503, and 967 sentences for informal, polite, and formal classes, respectively.)

3.2 Model

Drawing from the success of classifiers based on BERT (Devlin et al., 2019), which have achieved excellent performance in a large selection of downstream tasks, and following Liu and Kobayashi (2022), we propose to use Japanese-BERT (Suzuki and Takahashi, 2021) to train a formality classifier for Japanese formality. The input text is preprocessed and tokenized using the MeCab morphological parser (Kudo, 2005), which is what Japanese-BERT utilizes. For training, we used the AdamW (Loshchilov and Hutter, 2019) optimizer, with a learning rate of 10^{-5}, a batch size of 16 and, and train for a maximum of 20 epochs.

We evaluate our model in the test portions of existing datasets, namely, COCOA-MT and KEICO.

[3]We recruited 30 native Japanese speakers within 20 and 30 years old. All annotators are currently undergraduate or graduate students of a university in Tokyo, Japan. Furthermore, these annotations were double-checked with the help of Japanese dictionaries by 3 additional native Japanese speakers who are graduate students of the same university.

For the former, our analysis reveals that out of 600 examples in the test set, only 594 have been made available, which we utilize in our study. For the latter, since no official test splits are provided, we try to follow the experimental setting proposed by Liu and Kobayashi (2022)[4] and randomly selected 20% of the examples to test.

To contextualize our contributions and put the performance of our classifier in context, we consider a selection of baselines taken from previous work, as well as our implementations of newly-introduced models, and use F1-Score for evaluation.

On CoCoA-MT, we consider the rule-based classifiers proposed by Nadejde et al. (2022) and Feely et al. (2019), as well as our Transformer-based classifier. To test our model in this dataset, which is a binary classification setting, we either train another Transformer-based model on binary labels, or simply convert the prediction of the 3-way models into binary classification by considering all the other 3 tenses except for regular form as formal.

For this dataset, we additionally propose a new rule-based classifier for Japanese formality, which we adapt from Feely et al. (2019). Concretely, we propose ways to mitigate some limitations that were identified in the existing model. For example, the original rules assigned "ない (negative - present tense)" and "なかった (negative - past tense)" to the polite class, while we consider that both of them should be the regular form. In our approach, we label all sentences that are not classified as belonging to the "Polite" and the "Formal" class as "Informal".

We omit results by Rippeth et al. (2022), who fine-tune XLM-R on binary classification between formal and informal classes, but only report accuracy on the development set, defined as the last 50 paired contrastive examples from each language, which we regard as too small and incompatible with our setting. Their model obtains an accuracy of 98% on both the formal and informal classes on this set.

For the KEICO dataset, we compare the performance of our Transformer-based and rule-based classifiers against the BERT-based classifier proposed by Liu and Kobayashi (2022). Since this classifier is trained on a different label set compared to our approach, we proceed as follows: (1) we compare the average F1-score for the respectful and humble term detection task in (Liu and Kobayashi, 2022) against the F1-score of our classifier on the formal label, which we regard as a roughly-equivalent setting, (2) as our approach directly collapses Levels 1 and 2 in Liu and Kobayashi (2022) to our formal label, while Level 3 (which uses teineigo) and Level 4 (no honorifics) perfectly match our polite and informal class, respectively, we compare F1-scores as-is against the overall classification performance.

[4]Their reported metrics are the result of 10 runs with different initialization, and each time 20% of the examples are randomly chosen for the evaluation.

Model	Precision	Recall	F1-score
Nadejde et al. (2022)	0.70	0.49	-
Feely et al. (2019)*	0.87	0.83	0.83
Rule-based*	**0.98**	**0.98**	**0.98**
no KEICO samples			
Transformer 3-way*	0.97	0.97	0.97
Transformer 2-way	0.97	0.97	0.97
with KEICO samples			
Transformer 3-way*	0.97	0.97	0.97
Transformer 2-way	0.97	0.97	0.97

Table 1: Performance formality-level classifiers for Japanese on CoCoA-MT, where * indicates models that were originally designed for 3-way classification, but adapted for binary formality labels by considering polite, respectful, and humble forms as Formal. Precision and recall values from (Nadejde et al., 2022) are based on M-Acc score, and are computed on a 300-example subset of the data. F1-scores were not reported, so we omit them.

3.3 Results

As can be seen in Table 1, both our rule-based and Transformer-based models are able to outperform previous work on CoCoA-MT by substantial margins. We further notice that both models are able to attain very similar, and extremely high performance of 97% F1-score, and that neither the change in label setting, nor the addition of examples from KEICO have any effect on the performance. We think these results are compelling evidence suggesting the limited quality of the examples in this dataset. Based on this, we recommend researchers to consider other benchmarks instead.

Table 2 shows our results on the KEICO dataset. We see that our Transformer-based classifier obtains an overall F1-score of 0.84, surpassing of the classifier proposed by Liu and Kobayashi (2022). By contrast, our rule-based classifier only obtains

Model	F1-score	
	Formality	Hon. Level
Liu and Kobayashi (2022)	0.802	0.727
Rule-based	0.620	-
no KEICO samples Transformer	0.550	0.604
with KEICO samples Transformer	**0.840**	**0.810**

Table 2: Summary of results on the KeiCO dataset. The "Formal" column refers to the accuracy of the model to detect formal terms, while the "Level" column indicates the performance of detecting the level of honorifics.

an F1-score of 0.620, showing that rule-based methods, as comprehensive as they may be, offer limited reliability in multi-domain scenarios.

We also observe that the addition of examples from the KEICO dataset to the training data has a substantial impact on the performance of our model. It is only when these examples are added that our Transformer-based model is able to outperform the baseline. We think this result validates our domain-shift hypothesis, suggesting that examples from KOKAI offer only a narrow variety of expressions of Japanese formality, which do not allow the model to generalize well to more general domains.

Overall, our results suggest that the KEICO dataset offers a more compelling and real-like arena to evaluate the accuracy of Japanese formality classifiers.

4 Empirical Study

Having demonstrated the abilities of our Transformer-based classifier of Japanese formality, we now turn to a more practical issue, and tasks ourselves with testing the proposed approach in a real scenario. We examine formality abilities of English to Japanese machine translation using a zero-shot prompting approach. To the best of our knowledge, our work is the first one to study this issue.

4.1 Experimental Setup

Data We utilize the COCOA-MT En→Ja test set for our experiments. As mentioned earlier, examples in this dataset exhibit numerous flaws, including incomplete and semantically meaningless sentences, but since no other suitable dataset exists, we resort to this dataset nonetheless. We assume the existence of tuples $(x, y_{\text{formal}}, y_{\text{informal}})$ where x is the input sentence in English, and y are the target sentences in Japanese at different formality levels. Using the original 594 English sentences, below we show how we prompt our selection of models to produce both informal and formal Japanese translations.

Models We utilize large multilingual MT models trained on massive parallel corpora, specifically, M2M100 (Fan et al., 2021) and NLLB200 (Costa-jussà et al., 2022). Additionally, we experiment with M2M100 models of different sizes, including the 418M and 1.2B models. For each MT model, we use the English sentences from COCOA-MT as input, and concatenate them with a prefix prompt $p \in P = \{\text{formal}, \text{informal}\}$ which is added using square brackets. Thus, the input to the models is expressed as "$[p]; x$," where ; denotes white-space-based concatenation.

Moreover, as LLMs have shown good performance on MT when provided with an appropriate prompt, we also experiment using GPT-3 (Brown et al., 2020) and ChatGPT. We use similar prompts to those used for the MT models, but suggest more clearly to the models to perform the formality control task by using "Translate English to p Japanese: x". For ChatGPT, as the official API was not yet available at the time of our experiments, so we manually input a total of 1,188 examples (594 examples for each informal and formal setup) into the web client of ChatGPT Plus[5]. We also consider the recently-released llama2 models (Touvron et al., 2023), specifically the chat versions, which have been optimized for dialogue. We utilize the 7B-parameter and 13-B models, the latter we quantize to 4-bits using QLoRA (Dettmers et al., 2023) in order to fit our GPU memory. We follow the approach by the original paper to create our prompt, and test two settings (1) a zero-shot approach where the model is directly asked to generate translation, and (2) a one-shot setting, where we incorporate a source-target translation example for the given formality target. We construct this example manually, making sure it has minimum overlap with the examples from our data.

Finally, we also consider the Transformer-based model by Nadejde et al. (2022) as a baseline. This model is a 20-layer encoder and 2-layer decoder Transformer trained from scratch on the COCOA-MT, with the help of data augmentation techniques.

[5] https://openai.com/blog/chatgpt

Model	Cmp (%)	COMET	BLEU	M-Acc	Accuracy		
					Rule	T-3	T-2
Nadejde et al. (2022)	100	-	22.20	0.76 (-)	-	-	-
M2M100 (418M)	100	0.731	16.19	0.49 (0.18)	0.51	0.49	0.51
M2M100 (1.2B)	100	0.744	17.25	0.50 (0.19)	0.50	0.49	0.49
NLLB200 (600M)	100	0.733	8.83	0.47 (0.19)	0.49	0.48	0.48
llama2-chat (7B)	74.8	0.698	8.53	0.52 (0.24)	0.55	0.54	0.54
+ one shot	84.1	0.617	6.76	0.51 (0.26)	0.56	0.51	0.56
llama2-chat (13B)	55.4	0.731	11.36	0.83 (0.19)	0.64	0.63	0.63
+ one-shot	91.3	0.561	8.05	0.61 (0.30)	0.56	0.56	0.57
GPT-3	100	0.875	**23.79**	**0.86 (0.25)**	**0.91**	0.90	0.90
ChatGPT	98.9	**0.868**	20.63	0.83 (0.25)	**0.91**	**0.91**	**0.91**

Table 3: Performance of our experiments with formality-controlled En→Ja MT, including results MT models (Nadejde et al., 2022) fine-tuned on the data, and zero-shot approaches using pre-trained MT models and LLMs. Here, T-3 and T-2 indicate the proposed 3-way and binary Transformer-based classifiers, and Cmp. is short for Compliance, showing the percentage of output that contained valid translations. For M-Acc (Nadejde et al., 2022), we also show the coverage of the matched sentences between parenthesis, as this evaluation metric model overlooks examples that do not match its rules.

Evaluation We perform evaluation in terms of the quality of the generated translations, and in terms of the ability to perform formality control. For the former, we follow previous work and report and BLEU scores (Papineni et al., 2002) relying on the sacrebleu[6] implementation (Post, 2018), and also consider COMET (Rei et al., 2020b) using the "wmt22-comet-da" model, which has multilingual support. For the latter, we rely on Matched-Accuracy (M-Acc) (Nadejde et al., 2022) which is a rule-based corpus-level metric for CoCoA-MT. M-acc works by checking if the hypothesis contains: a) any of the formality-marking phrases annotated in the formal reference and b) none of the phrases annotated in the informal reference (or vice versa). Crucially, sentences that are not matched are simply skipped. This metric was shown by Nadejde et al. (2022) to be relatively reliable for Japanese, obtaining a precision and recall of 0.7 and 0.49, respectively, when tested on a random sample of 300 sentences that were manually annotated by two professional translators. Finally, we utilize our proposed rule-based and Transformer-based classifiers. Finally, we also measure the zero-shot or few-shot ability of LLMs to "comply" with the given prompt by generating plausible translations. Based on the provided instruction, we use heuristics to parse and extract the translation from the text generated, and report the percentage of output that our heuristics are able to parse successfully.

4.2 Results

Table 3 summarizes our results on the formality control in En→Ja MT performance of all the models considered. We see that zero-shot prompting techniques work much better on LLMs than on pre-trained multilingual MT models, with the former attaining the best performance overall. In particular, we see that zero-shot techniques based on prompting lead to substantially low BLEU scores and formality control accuracy when tested on pre-trained multilingual MT models, which are also outperformed by the fine-tuned models by (Nadejde et al., 2022). This suggests that pre-trained multilingual MT models may simply lack the ability to be prompted for formality control.

In terms of model compliance, we notice that prompting LLMs leads to unstable behavior, with models often not following the provided instruction. This therefore leads them to not generate a valid translation, or to do so in a what such that it is not feasible to find the translation automatically in the model output. For example, some models do not follow the input-output pattern described in the prompt, while others tend to explain their translations in some cases. Finally, llama2 models sometimes refused to provide a translation for

[6] https://github.com/mjpost/sacrebleu

safety reasons, as they detected words regarded as rude or potentially harmful in the input.

Moreover, our results shed light on the reliability issues of M-Acc which, due to its hard matching approach, ends up ignoring many of the translations generated by the systems we test. We observe that across all our tested systems, its coverage lies between 0.18 to 0.25. M-acc is in principle designed to work only for the CoCoA-MT dataset. While this is allegedly a strong limitation already, we think the coverage issue we observed suggests that the approach may be even more limited.

In contrast to these results, we observe that both our Transformer-based and rule-based approaches offer no coverage issues, while also agreeing with each other and with the overall M-acc scores. We think these results validate our techniques as valid alternatives for the evaluation of formality-controlled MT, setting a potential direction for future developments.

5 Conclusions

This paper explores new alternatives to evaluate the ability of En→Ja MT models to perform formality control, proposing classifiers based on rule-based methods and a machine learning approach using HuggingFace Transformers[7] (Wolf et al., 2020).

To build robust models, we focus on developing resources to improve formality detection in Japanese, uncovering several flaws on existing corpora for the task, and introducing new annotated datasets. In contrast to prior work approaching formality using binary labels, we use three classes (informal, polite, and formal) to better approximate the ways honorifics are used in the Japanese language. Extensive experiments on benchmark datasets show that our proposed models offer state-of-the-art performance.

Finally, we empirically show that our machine-learning approach is superior to existing evaluation techniques for formality-controlled MT systems, offering a reliable and accurate evaluation solution. The study also demonstrates the ability of LLMs to generate sequences with varying levels of formality through well-designed prompts, resulting in state-of-the-art results in En→Ja formality-controlled MT. Our findings provide a valuable contribution to the NLP field by presenting a new approach to evaluate formality-controlled MT systems and highlighting the effectiveness of LLMs in this task.

[7] https://huggingface.co/docs/transformers

Limitations

In this work, we have introduced both data and models to tackle the task of formality detection in the Japanese language. Though our results suggest that we have been able to build a robust classifier that obtains good performance, we offer no empirical evidence to suggest how well these capabilities could generalize to untested domains or situations.

Moreover, as some of our experiments involved black-box models that are only accessible through an API, such as GPT-3 and ChatGPT, we are unable to offer reliability in replicating those results. Upon acceptance, we will be releasing the output we obtained from these models for the sake of reproducibility of our experiments.

Finally, we also utilize pre-trained models either as baselines or to initialize our proposed classifier, and we think this is an important driver of the performance we observed. This may be an issue where access to pre-trained models is limited.

References

Antonios Anastasopoulos, Loïc Barrault, Luisa Bentivogli, Marcely Zanon Boito, Ondřej Bojar, Roldano Cattoni, Anna Currey, Georgiana Dinu, Kevin Duh, Maha Elbayad, Clara Emmanuel, Yannick Estève, Marcello Federico, Christian Federmann, Souhir Gahbiche, Hongyu Gong, Roman Grundkiewicz, Barry Haddow, Benjamin Hsu, Dávid Javorský, Věra Kloudová, Surafel Lakew, Xutai Ma, Prashant Mathur, Paul McNamee, Kenton Murray, Maria Nădejde, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, John Ortega, Juan Pino, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Yogesh Virkar, Alexander Waibel, Changhan Wang, and Shinji Watanabe. 2022. Findings of the IWSLT 2022 evaluation campaign. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 98–157, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Tamotsu Aoki, Yoshimitsu Aoyama, Takashi Atoda, Yoshiaki Ishizawa, Danjuro Ichikawa, Emi Uehara, Fumiko Okada, Tadaaki Odaka, Tsuneaki Kawamura, Yasuko Tabata, Takako Tamura, Hideki Tomizawa, Nakayama Nobuhiro, Kazuo Nishi, Suzuko Nishihara, Toyohiro Nomura, Tomihi Maeda, Kazuko Matsuoka, Mayumi Mori, and Nobuo Monya. 2007. Instruction to Japanese Formality (敬語の指針). Technical report, Agency for Cultural Affairs, Government of Japan (文化審議会).

Douglas Biber and Susan Conrad. 2019. *Register, genre, and style*. Cambridge University Press.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond English-Centric Multilingual Machine Translation. *Journal of Machine Learning Research*, 22(107):1–48.

Weston Feely, Eva Hasler, and Adrià de Gispert. 2019. Controlling Japanese honorifics in English-to-Japanese neural machine translation. In *Proceedings of the 6th Workshop on Asian Translation*, pages 45–53, Hong Kong, China. Association for Computational Linguistics.

Atsushi Fukada and Noriko Asato. 2004. Universal politeness theory: application to the use of Japanese honorifics. *Journal of Pragmatics*, 36(11):1991–2002.

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10867–10878. PMLR.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Takumitsu Kudo. 2005. Mecab : Yet another part-of-speech and morphological analyzer. *http://mecab.sourceforge.jp*.

Muxuan Liu and Ichiro Kobayashi. 2022. Construction and validation of a Japanese honorific corpus based on systemic functional linguistics. In *Proceedings of the Workshop on Dataset Creation for Lower-Resourced Languages within the 13th Language Resources and Evaluation Conference*, pages 19–26, Marseille, France. European Language Resources Association.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. 2018. *Foundations of Machine Learning*, 2 edition. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA.

Maria Nadejde, Anna Currey, Benjamin Hsu, Xing Niu, Marcello Federico, and Georgiana Dinu. 2022. CoCoA-MT: A dataset and benchmark for contrastive controlled MT with application to formality. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 616–632, Seattle, United States. Association for Computational Linguistics.

Xing Niu and Marine Carpuat. 2020. Controlling neural machine translation formality with synthetic supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8568–8575.

Xing Niu, Marianna Martindale, and Marine Carpuat. 2017. A study of style in machine translation: Controlling the formality of machine translation output.

In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2814–2819, Copenhagen, Denmark. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Dongqi Pu and Vera Demberg. 2023. ChatGPT vs human-authored text: Insights into controllable text summarization and sentence style transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1–18, Toronto, Canada. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020a. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020b. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Elijah Rippeth, Sweta Agrawal, and Marine Carpuat. 2022. Controlling translation formality using pre-trained multilingual language models. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 327–340, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *International Conference on Learning Representations*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, San Diego, California. Association for Computational Linguistics.

Masatoshi Suzuki and Ryo Takahashi. 2021. Japanese bert. `https://github.com/cl-tohoku/bert-japanese`.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

There's no Data Like Better Data: Using QE Metrics for MT Data Filtering

Jan-Thorsten Peter* **David Vilar*** **Daniel Deutsch**
Mara Finkelstein **Juraj Juraska** **Markus Freitag**
Google
{jtp,vilar}@google.com

Abstract

Quality Estimation (QE), the evaluation of machine translation output without the need of explicit references, has seen big improvements in the last years with the use of neural metrics. In this paper we analyze the viability of using QE metrics for filtering out bad quality sentence pairs in the training data of neural machine translation systems (NMT). While most corpus filtering methods are focused on detecting noisy examples in collections of texts, usually huge amounts of web crawled data, QE models are trained to discriminate more fine-grained quality differences. We show that by selecting the highest quality sentence pairs in the training data, we can improve translation quality while reducing the training size by half. We also provide a detailed analysis of the filtering results, which highlights the differences between both approaches.

1 Introduction

In the times of statistical machine translation, a well-known motto was "there's no data like more data". Experimental results seemed to confirm this, with the performance of the systems steadily improving as more data was made available. Web crawling has proven to be a valuable source of data for training translation systems, with projects like Common Crawl[1] or ParaCrawl (Bañón et al., 2020) providing numerous parallel sentences with which to train MT systems. Inevitably, when crawling huge amounts of data, noise will be present. Taking the web as an example, the quality of available texts varies greatly between websites. There are sources which reliably produce high-quality text, e.g. large circulations newspaper websites usually contain text written and proof-read by professional journalists. But by its open nature, the web also contains texts of dubious quality (both in style and in content) which may pollute the collected texts.

When considering bilingual data collection, an additional difficulty comes into play, namely the alignment of segments between two or more languages. Sentence alignment algorithms (Gale and Church, 1993; Moore, 2002; Sennrich and Volk, 2011; Thompson and Koehn, 2019) are bound to make mistakes, resulting in pairing of sentences that are not necessarily translations of each other. Even if the correspondence between sentences may be correct, the quality of sentences may differ greatly between languages. While one source may provide high quality text in its original language, the available translations may be of sub-par quality due to a variety of reasons (Freitag et al., 2022b). In addition, given the increased availability (and quality) of machine translation engines, MT output is expected to be part of the crawled data, thus contaminating the training material.

Statistical systems were robust against such type of noise (Goutte et al., 2012). The maximum likelihood estimators of phrase probabilities (and related models) were based on relative frequencies, with the consequence that noisy translation units, while being available to the system at translation time, had a low chance of being used. In fact, works on filtering data dealing with statistical systems, e.g. Johnson et al. (2007) were more concerned with the efficiency of the systems, rather than with quality.

With the advent of neural machine translation, the situation has changed, and the quality of the data has a major impact in the resulting quality of the translation system. Neural networks have a great ability of memorizing (parts of) the training data (Arpit et al., 2017; Feldman and Zhang, 2020). Whereas for phrase-based models the noise was diluted in the abundance of better-quality data, in neural models such outliers may have a critical effect on the output of the system. Therefore data filtering has become increasingly important, even spawning dedicated shared tasks (Koehn et al., 2018, 2019,

* Equal contribution.
[1] https://commoncrawl.org/

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 561–577
December 6–7, 2023.

2020). Research in this area has allowed NMT systems to take advantage of big amounts of training data. As an example, initial versions of ParaCrawl degraded translation quality when adding them unfiltered to the training data of MT systems (Junczys-Dowmunt, 2018b; Schamper et al., 2018), whereas nowadays it is one of the main data sources used in WMT for the languages where it is available.

Thanks to these filtering techniques, huge amounts of parallel sentences are available for several languages, in some cases reaching up to hundreds of millions of sentences. But as noted above, not all texts are of the same quality. It is only natural to ask the question, once the training data reaches a size which is "big enough", if all the available text is useful for training NMT systems, or if the lower quality sentences are hindering the system. Note that most data filtering systems are focused in detecting noise or problems in the translation (e.g. under- or overtranslation). In very wide strokes, most systems answer the question "Is sentence A a translation of sentence B?", without looking (too much) into the quality.

Judging the quality of translations is the focus of the Quality Estimation (QE) field of machine translation. It can be considered an extension of machine translation evaluation, where references are not available. In the last years, the use of neural models has improved the results in this area dramatically. Would it then be possible to use quality estimation methods for filtering data and improve the quality of a neural machine translation system, which has been trained on already cleaned data? In other words, can we do a more fine-grained data selection beyond discarding "obvious" errors, focusing on selecting the best data that can be found in the training corpus? We explore these questions in this paper.

Our scientific contributions are:

- We show that neural QE metrics are effective methods for data filtering.
- We analyze the differences between the sentences filtered by the methods and find out that QE methods are more sensitive toward quality differences, being able to detect bad quality translations or fine grained translation errors (e.g. wrong named entities in a perfectly valid translation).
- We show that on the other hand QE filtering cannot account for some actual noise problems. Thus a "traditional" method for filtering the raw data coming from crawls is still needed as a first step.

2 Related Work

Already in the early days of the popularization of statistical methods for machine translation, the potential of mining data from the web was recognized by Resnik (1999). In contrast to other data sources at the time (e.g. Canadian Hansards, European Parliament Proceedings) which consisted largely of clean data, the necessity of including additional "quality assurance" steps were recognized in this work. As pointed out above, statistical systems were robust against noisy input data (Goutte et al., 2012), and as such, the topic of "corpus filtering" was mainly focused on selecting subsets of data closer to a given domain (Axelrod et al., 2011). Nevertheless Taghipour et al. (2011) shows that statistical systems may also benefit from careful curation of the training data.

The situation changed dramatically with the advent of neural machine translation, as such systems are much more sensitive to noisy input data (Khayrallah and Koehn, 2018). A clear reflection of this fact was the creation of a new dedicated shared task in the WMT yearly conference (Koehn et al., 2018, 2019, 2020) in the years 2018 to 2020.[2]

Junczys-Dowmunt (2018a) was the best performing system in the first edition of the filtering shared task, using a cross-entropy approach between two translation systems trained on clean data. In the next edition, the focus was moved towards low resource conditions. That year Chaudhary et al. (2019) presented the best performing system, using a system based on LASER embeddings (Schwenk and Douze, 2017).

The 2020 edition continued the focus on low-resource languages. At that evaluation, three were the best performing systems: Lu et al. (2020) and Lo and Joanis (2020) both use pre-trained multilingual models as a key component of their filtering systems. Esplà-Gomis et al. (2020) used an improved version of BICLEANER, their submission for the 2018 campaign (Sánchez-Cartagena et al., 2018). The authors further improved their system (Ramírez-Sánchez et al., 2020), including an extension using neural models (Zaragoza-Bernabeu et al., 2022). This latest version is considered state-

[2] In this year's WMT there is a new related shared task: "Parallel Data Curation".

of-the art and it is that which we take as baseline for comparing our method.

Of course, this is just but a very rough overview of the best performing systems in each evaluation campaign. We refer to the reports of each campaign for a more detailed overview of the methods explored each year. Bane et al. (2022) provide one more recent overview of data filtering methods. In this work, the authors sample 5M sentences from original training data and added 1M noise samples manually. They show that a two-stage approach can be beneficial for improving the quality of a translation system.

All of these methods have a common focus on detecting the type of noise that may originate from crawled data. This type of noise has been analyzed in Khayrallah and Koehn (2018), and Herold et al. (2022) build on their work and carry out a comparison of the efficiency of different filtering methods on various types of noise. Kreutzer et al. (2022) also provide an extensive analysis on the noise present in several widely used corpora.

Most similar to our work Carpuat et al. (2017) start with already clean data and analyze the effect of semantic divergence on translation quality. They are able to effectively select a subset of the training data and improve translation quality measured in BLEU. Bernier-Colborne and Lo (2019) use YiSi-2, also a quality estimation metric as a component in their corpus filtering system for the WMT 2019 shared task. Lo and Simard (2019) extend this idea by including BERT (word) alignments in the YiSi pipeline. We follow a conceptionally similar approach to these papers, using state-of-the-art QE metrics and provide a more in-depth comparison to other corpus filtering methods more oriented towards noise detection.

Quality estimation is again its own area of research, with dedicated shared tasks, e.g. (Zerva et al., 2022), that measure how well metrics can predict word- and sentence-level quality scores. In contrast to traditional MT evaluation, QE aims to assess the quality of the output texts without the use of a reference translations. The most successful QE metrics learn to jointly predict word- and sentence-level scores, like COMETKIWI (Kepler et al., 2019; Rei et al., 2022). Another possibility is to modify the input to a learned reference-based metrics like BLEURT (Sellam et al., 2020) or COMET (Rei et al., 2020) to use the source segment instead of a reference translation to predict sentence-level quality scores (Rei et al., 2021). We follow the latter approach and train a QE version of BLEURT that predicts sentence-level quality scores (see Section 3) that are used for data filtering.

3 From BLEURT to BLEURTQE

The QE metric that we propose for data filtering is a learned MT evaluation metric that is based on a BLEURT-style architecture (Sellam et al., 2020). BLEURT is a reference-based regression metric that is trained to predict a quality score for a hypothesis translation given a reference. The hypothesis and reference are concatenated together with a special token in between, then fed as input to the metric, which predicts a floating point quality score.

Our QE metric is a modification of the original model. To make it a QE metric, we pass the source segment as input to the metric instead of the reference. Then, we follow the winning submission to the WMT'22 Metrics Shared Task (Freitag et al., 2022a), MetricX, and use a modified version of the mT5 encoder-decoder language model (Xue et al., 2021) as our network architecture. Not that these is a multilingual model, so the same system can be used for a variety of languages. The source and hypothesis are passed as input to the encoder, and an arbitrary logit from the first step of the decoder is trained to predict the hypothesis quality score.

The QE metric is trained on the direct assessment quality judgments that were collected as part of the WMT Metrics Shared Task from 2015-2020 (Bojar et al., 2015, 2016, 2017; Specia et al., 2018, 2020; Fonseca et al., 2019) for all available language pairs. To (meta-)evaluate the metric we measure its correlation with ground-truth translation quality ratings using the benchmark MQM dataset from WMT'22 (Zerva et al., 2022) that includes 3 language pairs: en-de, zh-en, and en-ru. Since our metric is used to score individual segments and not systems, we report the segment-level correlation between our metrics' scores and the gold MQM scores using Pearson's r and Kendall's τ, shown in Table 1. The correlations are competitive to the top QE submissions to the WMT'22 Metrics Shared Task.

A (more refined) version of this metric has been submitted to this year's QE shared task (Juraska et al., 2023), and has been open sourced. We refer the reader to the system description for a more fine-grained discussion of the details of the metric.

Metric	en-de		en-ru		zh-en	
	r	τ	r	τ	r	τ
UniTE-src	0.40	0.29	0.39	0.34	0.40	0.43
COMETKIWI	0.43	0.29	0.39	0.36	0.51	0.36
BLEURTQE	0.38	0.29	0.41	0.39	0.38	0.35

Table 1: Segment-level Pearson's r and Kendall's τ on the WMT'22 MQM ratings for our QE metric, BLEURTQE and the top-performing metrics in the WMT'22 Metrics Shared Task, COMETKIWI (Rei et al., 2022), UniTE-src (Wan et al., 2022).

4 Experiments

We report experiments on three language pairs: English ↔ German, Japanese ↔ English and Chinese ↔ English. Our starting point is the full training data as provided by the WMT evaluation campaign. Corpus sizes can be found in Table 3. As can be seen in that table, we are working on a medium-to-large data condition, with the smallest language pair already having over 30M sentence pairs.

One thing to note is that these datasets have already undergone a cleaning process by the WMT organizers. I.e. a system trained on the entirety of this data is already able to obtain very good performance. In fact, many of the systems participating in the WMT evaluations take the available data as-is.

For each language pair we will consider different ways to reduce the size to 50% of their original size. This value was chosen in preliminary experiments on the English to German data, and it is comparable to previous work (Bane et al., 2022). Fixing the target size beforehand also allows a fair comparison between all the methods.

4.1 Filtering Approaches

We will consider three different filtering approaches for our experiments.

4.1.1 Random Selection

The most straightforward method to reduce the size of the training data is to just randomly select the desired amount of sentence pairs. We do not expect this method to perform well, but it constitutes the most direct baseline for data size reduction.

4.1.2 BICLEANER

As a representative for the "noise-detection" corpus filtering methods we chose to use BICLEANER AI.[3] This tool is an extension of the previous BICLEANER tool. The underlying method is based on a classifier that predicts if a sentence is a translation of another. BICLEANER AI substitutes the original classifier, based on handcrafted rules and extremely randomized trees, with a neural classifier based on XLM-RoBERTa. Zaragoza-Bernabeu et al. (2022) provide a detailed description of the tool and present an extensive experimental comparison showing state of the art results for filtering ParaCrawl.

It is also worth noting that BICLEANER is part of the pre-processing pipeline for generating the ParaCrawl dataset.

4.1.3 Quality Estimation for Filtering

For testing the performance of QE metrics for filtering we use two state-of-the-art metrics, COMETKIWI[4] (Kepler et al., 2019) and BLEURTQE[5] as described in Section 3. For each sentence pair in the training data, we compute the QE score for the translation from English into the foreign language. We use these scores for filtering for both translation directions, i.e. the resulting parallel data is the same for English → Foreign than from Foreign → English. We are aware that this may introduce a certain bias, as the performance of the QE metrics is not symmetrical. However scoring the full training data is a costly operation as we have to run big neural models on tens or hundreds of millions of sentence pairs. We still expect to see improvements even when using the wrong direction for data filtering. The only exception may be the backtranslated portion of the Chinese ↔ English dataset: As the starting data is Chinese, the filtering method may miss low quality backtranslations produced by an automatic system.

4.2 Experimental Setup

For all the filtering methods (except random selection), we compute the score of each sentence pair, and then select a threshold as to keep 50% of the original data. We then train an NMT system from scratch using the resulting training data sets.

Our translation system is a transformer-based encoder-decoder model based on PaxML[6], very similar to most of the systems participating in the WMT evaluation campaign. It consists of 6 encoder

[3] https://github.com/bitextor/bicleaner-ai

[4] https://unbabel.github.io/OpenKiwi

[5] The tool will be open-sourced with the publication of the shared task system description.

[6] https://github.com/google/paxml

and 6 decoder layers, a model dimension of 1024, hidden dimension of 8192 and 16 attention heads. GELUs with gated activation are used as activation functions. We use a 32k shared vocabulary for each language pair, and limit the maximal sentence length to 128 tokens. The model has a total of 551M parameters.

We removed all sentences which have more than 128 tokens, but did not perform any other filtering or preprocessing of the data. All models are trained until they converged and we selected the checkpoint with the best BLEURT score on the WMT 2022 test sets.

In the discussion of the results we focus on the evaluation using COMET22. Traditional metrics like BLEU and CHRF are consistently outperformed by neural metrics in the WMT metrics shared task (Freitag et al., 2022a), thus we favor the use of such new metrics. We chose COMET22 over BLEURT in order to avoid overfitting on this last metric, as our proposed BLEURTQE model is based on it, and it also guides the checkpoint selection. Nevertheless, BLEURT, BLEU and CHRF scores are given in Appendix B.2 and confirm the trends reported here.

4.3 Test Data

In order to test on a variety of domains we use test sets from the WMT and IWSLT evaluation campaigns. We use the WMT 2019 (where available) consisting of news data, and the WMT 2022 and WMT2023 test sets, which are composed of a mix of different domains each. Additionally we experiment on the IWSLT'21 test set, sourced from TED talks (Anastasopoulos et al., 2021), and the IWSLT'23 dev set[7], which is based on ACL talks presentations (Agarwal et al., 2023).

Following the training data settings, we also filtered the test sentences longer than 128 tokens. As the WMT 2023 test set includes paragraph level evaluation, its size is reduced for en → de from 557 segments to 404 and for de → en from 549 to 468. All other test sets are barely affected (see Table 6 in Appendix A).

4.4 Experimental Results

Translation results for the English ↔ German language pair are shown in Table 2a. For en → de we can see that randomly selecting data hurts performance by 1 point on the WMT23 test set. Using each of the other filtering methods we are able to improve performance over using the full training dataset. For BICLEANER the improvement is rather modest, around 0.4 points for most test corpora. Note however that BICLEANER was already applied to the ParaCrawl dataset, which constitutes a big portion of the available training data for this language pair. As such it is understandable, or even expected, that translation quality is not improved by applying it again. The QE metrics perform similarly to each other, with a slight advantage of BLEURTQE over COMETKIWI. Using BLEURTQE we are able to achieve an improvement of up to 1.7 points on the WMT23 test set.

The results for de → en majorly confirm the previous observations. The best results are again achieved in this case on the WMT'23 data, with an improvement of 1.3 points achieved by both QE methods. For the WMT'22, IWSLT'21 and IWSLT'23 test sets, the translation performance basically stays constant for all filtering methods.

Results on English ↔ Japanese, shown in Table 2b also show similar trends. In this case the biggest improvement comprises 2.3 points on the WMT'23 test set[8], obtained by BLEURTQE. However for the ja → en translation direction we find an outlier, where no filtering achieves improvements over the baseline on the IWSLT'23 data.

Lastly, Table 2c shows the results for the Chinese ↔ English language pair. Again we can confirm the same trends as for the other two language pairs. The QE metrics are able to improve up to 2.8 points for the WMT'23 test set. The IWSLT'23 dataset again fails to achieve improvements, and in this case BICLEANER deteriorates translation quality, while the QE metrics are able to keep the performance.

Overall, we see that the QE metrics are effective in improving translation quality while retaining just half of the training data. The improvements can rage up to more than 2 COMET22 points, depending on language pair and test set. For IWSLT'23, having a more specialized technical domain, the QE metrics are not able to improve quality, for several language directions. But except for the case of English → Japanese, they also do not hurt performance. Additional results differentiating between the single domains of the WMT'22 and WMT'23 corpora can be found in Appendix B.1. In Ap-

[7]We use the dev set for IWSLT'23, since the test set is currently not publicly available.

[8]A slightly bigger improvement of 2.4 is obtained for WMT'22, but we skip this as we used this corpus to choose the best checkpoint during training.

(a) COMET22 scores for en ↔ de experiments.

	Filter	WMT'22 (dev)	WMT'19	WMT'23	IWSLT'21	IWSLT'23
en → de	Random	84.0	85.5	80.8	82.8	84.2
	None	86.2	86.0	81.8	83.2	84.5
	BICLEANER	86.7⋆	86.4⋆	82.2	83.3	84.9
	COMETKIWI	86.9⋆⋆	**86.7**⋆⋆	82.9⋆	83.6⋆	84.8
	BLEURTQE	**87.2**⋆⋆⋆	**86.7**⋆⋆	**83.5**⋆⋆	**83.9**⋆⋆	**85.1**⋆
de → en	Random	84.2	84.3	83.5	84.1	87.2
	None	84.5	84.6	83.3	84.2	**87.4**
	BICLEANER	84.2⋆	84.8	84.0⋆	84.1	87.3
	COMETKIWI	84.6⋆	85.1⋆	**84.6**⋆⋆	**84.4**⋆	87.3
	BLEURTQE	**84.8**⋆⋆	**85.2**⋆	**84.6**⋆⋆	**84.4**⋆	87.3

(b) COMET22 scores for en ↔ ja experiments.

	Filter	WMT'22 (dev)	WMT'23	IWSLT'23
en → ja	Random	84.5	80.7	85.8
	None	85.6	82.3	86.9
	BICLEANER	86.0⋆	83.2⋆	87.2
	COMETKIWI	86.6⋆⋆	83.7⋆⋆	**87.9**⋆
	BLEURTQE	**87.0**⋆⋆⋆	**84.0**⋆⋆⋆	87.4
ja → en	Random	75.9	75.0	84.6
	None	77.6	75.9	**85.5**⋆
	BICLEANER	78.1⋆	77.4⋆	85.0
	COMETKIWI	78.7⋆⋆	78.0⋆⋆	85.0
	BLEURTQE	**79.0**⋆⋆⋆	**78.2**⋆⋆	85.1

(c) COMET22 scores for en ↔ zh experiments.

	Filter	WMT'22 (dev)	WMT'19	WMT'23	IWSLT'23
en → zh	Random	80.2	77.2	79.3	82.1
	None	81.2	77.6	79.7	**84.2**⋆
	BICLEANER	81.7⋆	78.4⋆	80.3⋆	83.3
	COMETKIWI	83.0⋆⋆	**80.1**⋆⋆	**82.5**⋆⋆⋆	84.1⋆
	BLEURTQE	**83.4**⋆⋆⋆	79.9⋆⋆	82.2⋆⋆	84.1⋆
zh → en	Random	72.2	78.1	74.2	84.1
	None	72.8	78.3	74.7	**84.9**⋆
	BICLEANER	74.8⋆	79.6⋆	75.7⋆	84.2
	COMETKIWI	75.2⋆⋆	**80.0**⋆⋆	**76.0**⋆⋆	84.5
	BLEURTQE	**75.4**⋆⋆	79.9⋆⋆	**76.0**⋆⋆	84.8⋆

Table 2: COMET22 scores all experiments. For each language direction, systems marked with stars are statistically significantly better than systems with fewer stars (pairwise permutation test (Koehn, 2004) with p=0.05). "Random" was excluded from the significance computation.

Language pair	Full	Filtered	Common
de ↔ en	292.8M	146M	105M
en ↔ zh	55.2M	27.6M	17.5M
en ↔ ja	33.9M	16.9M	10.4M

Table 3: Amount of sentences before filtering, after filtering, i.e. 50% of the original corpus size, and number of sentences kept by both BLEURTQE and BICLEANER. All language directions include ParaCrawl data. English ↔ Chinese includes around 19.7M backtranslated Chinese sentences, as provided by the WMT organizers.

pendix B.2 we also report the results of the same experiments using BLEU, CHRF and BLEURT. These metrics confirm the observations presented in this section.

Koehn et al. (2020) mentions that on average metrics that select shorter sentences performed better on Parallel Corpus Filtering. Contrary to that we observed that dropping 50% of the data with the proposed method led to on average slightly longer sentences e.g. from 14.4 to 15.5 words per sentence for BLEURTQE on English ↔ German.

5 Analysis

In this section we provide an in-depth analysis of the differences between the BLEURTQE-based and the BICLEANER filtering methods. Table 3 shows the amount of sentence pairs that are kept by both methods, which is roughly two thirds of the filtered sentences for all language pairs. Thus, it is clear that both methods do indeed perform quite different filtering. We will first report on manual inspection of the most striking divergences between both methods. In Section 5.2 we will then provide a more quantitative analysis of the behaviour of the methods using synthetic data.

5.1 Human Inspection

We will now analyze the difference in the filtering methods by looking into the sentences that are selected by each method. To this end, we select the sentences where one method filters it but the other does not. In addition we use automatic clustering methods in the spirit of (Aharoni and Goldberg, 2020) in order to get insights about topic distribution. We limit our analysis to the German–English language pair[9], but as the methods are largely language independent, we feel confident that our findings will generalize to the other language pairs. Also, due to the fuzzy and partially subjective nature of this investigation, we are unable to provide exact statistics about each kind of effect.[10]

We have encountered the following major differences in the working of the methods. For each of these categories it is easy to find an abundance of examples (easily in the thousands) in the filtered data.

Single Entity Mistranslations When looking into the parallel data available for training, one can find a big amount of "templated texts", i.e. sentences that have a common structure, but that differ in one or few components, frequently named entities or numbers. Some examples can be found in Table 4a. The first entry in this table is a typical example. In the travel domain, there is a big amount of sentences of the form "Flights from *cityA* to *cityB*", "Hotels in *city*" or similar formulations. One frequent source of sentence alignment errors originates from sentences that follow the same template, but have different instantiations. Although the travel domain is one of the biggest representative of these type of sentences, it is by no mean the only one, as the other examples in Table 4a show, including the financial and the technical domain.

In these type of sentences, the QE metric seems to be more sensitive to alignment errors. All the sentences shown in Table 4a (and many others) are selected by BICLEANER, while BLEURTQE discards them.

Low Quality Translations In this category we include training examples where one or both sides are of low quality. Examples can be found in Table 4b. We can see that the language quality of the examples is borderline at best. Strictly speaking, the translations are "correct" in the sense that they preserve the structure of the sentence. As such BICLEANER gives them a relative high score and are kept in the training corpus. BLEURTQE, on the other hand, is explicitly trained to flag such erroneous sentences (as they might very well originate from MT engines), and thus these examples are filtered out.

Bad Related Sentence Alignments As pointed out above, sentence alignment is also an automatic process. While both methods perform quite well when detecting clearly bad aligned sentences (see Section 5.2), we found that there are cases where

[9] None of the authors are speakers of Japanese or Chinese.

[10] If we were able to develop such statistics in an automatic way, we would be able to improve the filtering methods by including the same approaches!

(a) Single Entity Mistranslations. Templated texts where the specific instantiation is different in both languages. BLEURTQE filters out these examples, while BICLEANER keeps them.

English	German	Comments
Flights from Tallinn to Stockholm	Flüge ab Tallinn nach Friedrichshafen	"Stockholm"' changed to "Friedrichshafen".
Total EU spending in Germany – € 11.013 billion	Gesamtzuschüsse der EU in den Niederlanden: 2,359 Milliarden EUR	Land and amount changed.
Documents that we receive from a manufacturer of a Redball Electrical 565 can be divided into several groups.	Dokumente, die wir vom Produzenten des Geräts Trevi AVX 565 erhalten, können wir in mehrere Gruppen teilen.	Product code changed.

(b) Examples of low quality sentences in at least one of the languages. BLEURTQE filters out these examples, while BICLEANER keeps them.

English	German	Comments
We are both, we have own factory which can ensure sculpture quality and best price and have a profession team to provide you best service.	Wir sind beide, wir haben eigene Fabrik, die Skulpturqualität und besten Preis sichern kann und ein Berufsteam haben, um Ihnen besten Service zur Verfügung zu stellen.	Unnatural language on both sides.
We honor do not track signals and do not track, plant cookies, or use advertising when a Do Not Track (DNT) browser mechanism is in place.	Wir achten darauf, dass Sie keine Signale verfolgen und keine Cookies verfolgen oder Cookies verwenden, wenn Sie einen DNT-Browser-Mechanismus (Do not Track) verwenden.	Unnatural language on both sides.
It really is fast, easy, free and additionally to attempt.	Es ist schnell, Schnell, gratis und am besten von allen zu try.	Incorrect sentences in both languages.

(c) Examples of wrong sentence alignment, although the sentences are related to each other. BLEURTQE filters out these examples, while BICLEANER keeps them.

English	German	Translated German
Could you help me? Help to improve my English and French language	Ja, ich möchte gern mein Deutsch mit Dir verbessern.	Yes, I want to improve my German with you.
We, therefore, guarantee that you will get daily updates on office spaces to rent in Hong Kong.	Wir können Ihnen deshalb versichern, dass Sie bei uns täglich einen aktuellen Überblick über den österreichischen Markt erhalten.	We can guarantee that you will get an up-to-date daily overview about the Austrian market.
This implies that the law is either repealed or not enforced.	Darüber hinaus wird sichergestellt, dass bestehende Gesetze nicht dupliziert oder konterkariert werden.	In addition, it is ensured that existing laws are not duplicated or counteracted.

(d) Examples of sentence pairs originating from the Bible. BLEURTQE filters them out, probably due to archaic language, while BICLEANER keeps them.

English	German
19 Behold, my belly is as wine which hath no vent; it is ready to burst like new bottles.	19 Siehe, mein Bauch ist wie der Most, der zugestopfet ist, der die neuen Fässer zerreißet.
7:16 Those who went in, went in male and female of all flesh, as God commanded him; and Yahweh shut him in.	7:16 und das waren Männlein und Fräulein von allerlei Fleisch und gingen hinein, wie denn Gott ihm geboten hatte.

Table 4: Example sentences where the filtering methods diverge.

the source and target sides are related and the BICLEANER system seems to get confused by this proximity. Some examples are given in Table 4c. It can be seen that in all three examples the German side is clearly related to the English text, with probably a overlap big enough to get an acceptable score from the translation system underlying BICLEANER. Again, as BLEURTQE is trained to distinguish fine-grained differences between translations, it is more robust against this kind of problems.

Religious Texts One shortcoming we found for the BLEURTQE method is that many sentences originating from the Bible corpus (or similar religious texts) are filtered out, while BICLEANER keeps them. Some examples are given in Table 4d. This is probably due to the language being archaic, very different to the type of sentences BLEURTQE has been trained on. Such style would be heavily penalized in an evaluation, as a more modern language would be preferred.

5.2 Noise

In the previous section we saw several examples where BLEURTQE outperforms BICLEANER for data selection. However we should not forget that BICLEANER was developed with a (related but) different goal, namely the cleaning of raw data. In fact, our starting datasets, as made available for the WMT evaluation have already undergone a cleaning process, and are already at a pretty high quality level.

If we were dealing with crawled data directly, we would need to address different phenomena. In this section we study how the filtering methods perform when dealing with the typical noise found on crawled data. We follow Herold et al. (2022) for the categorization of different noise types, which in turn is based on Khayrallah and Koehn (2018). We create synthetic data for the English → German translation direction containing the following noise categories:

Misaligned Sentences created by shuffling the target side of the corpus.

Misordered Words created by reordering the words in either the source or the target sentences.

Wrong Language created by taking parallel sentences corresponding to another language pair.

Untranslated created by copying one sentence into the other direction i.e. each sentence pair in the corpus has a copy of the source sentence as a "target" sentence (or the reverse direction).

Over/Undertranslation created by truncating either the source or the target side.

We refer the reader to Herold et al. (2022) for a more detailed description and justification of these categories. We omitted the "Short Segments", "Raw Crawled Data" and "Synthetic Translations" categories, as it was not clear how to define the correct filtering strategy in those cases.

For each of the studied categories, we generated 200K synthetic noise examples by randomly selecting a subset of the training data. For these experiments we re-tuned the threshold for each method by computing the scores for the original sentences and the noise examples, and computing the median. In this way, we filter exactly half of the data and a perfect system would be able to completely separate the original examples from the noisy ones.

Results can be found in Table 5. It can be seen that for most categories BICLEANER clearly outperforms BLEURTQE. This is specially the case for the "Wrong Language" and "Untranslated" categories, where BICLEANER can detect all the noisy examples. In fact, one of the practical advantages of BLEURTQE is at the same time one of its weaknesses. As its backbone model is a multilingual model, it is able to handle a wide number of languages, but it does not have a way to differentiate between them.

For "Misordered Words" we find an interesting asymmetry. BLEURTQE is much stronger in detecting problems when the target side is reordered, undoubtedly due to this being the "natural" direction for which it was trained. BICLEANER also shows this behavior, with its target side performance being superior to that of BLEURTQE, but inferior in the opposite direction. BICLEANER is also clearly better at detecting Over- and Undertranslations.

5.3 Combination of Filtering Methods

Since BLEURTQE and BICLEANER based filtering both improve translation quality, and they filter different sentences, it is only natural to try to combine both. As can be seen in Table 3, the amount of available data dropped to roughly one third when combining both methods. The result

Noise type	BLEURTQE	BICLEANER
Misaligned Sentences	8.1	**5.6**
Misordered Words (src)	**24.3**	39.1
Misordered Words (tgt)	10.3	**6.3**
Wrong Language	43.8	**0.0**
Untranslated (src)	47.6	**0.0**
Untranslated (tgt)	63.8	**0.0**
Overtranslation	35.2	**13.9**
Undertranslation	13.4	**7.5**

Table 5: Percentage of sentences being kept as valid for each of the synthetic noise categories. 0% means that all noisy sentences have been filtered out, i.e. perfect performance.

only slightly degraded in the English to German direction compared to just using BLEURTQE (0.3 BLEURT points on WMT'22), but degraded more in the German to English direction (0.6 BLEURT points on WMT'22). Combining the two methods is thus too aggressive with our setup, and hurts translation performance. Adapting the thresholds for the combination, may result in better performance. Note however that ParaCrawl already used BICLEANER in its pipeline (Esplà et al., 2019), thus we have already implicitly been using a combination of both methods.

6 Conclusions

In this paper we have shown that filtering data using QE metrics is an effective way of improving translation quality. In contrast to "traditional" data filtering methods that focus on detecting noise in the data, QE methods focus on selecting the best translation examples. Analyzing the differences between the two different methods, we see that QE metrics are not as effective at detecting certain types of noise, e.g. untranslated sentences, but are much better at identifying more fine grained problems in the data, like small translation errors or grammatical mistakes. Therefore, when starting with already cleaned data, we can obtain a boost in performance by focusing the NMT system training on the best sentences.

Our results show that the improvements obtained generalize across different domains, as measured by a variety of metrics. Even for more distant domains, like the ACL Talks of of the IWSLT'23 corpus, the performance of the systems remains largely constant. QE estimation is a very active field of research. Using this approach, the improvements obtained in this area can have a direct impact on improving the quality of NMT systems.

Limitations

Better results could have been obtained by tuning the threshold for each method individually, but this would also increase the computational cost massively.

A more in-depth comparison could be carried out starting from the raw web-crawled data. However in this study we chose to start from conditions similar to what most participants in the WMT evaluation use.

Ethics Statement

BLEURTQE and COMETKIWI scoring all the training data is computationally expensive, and may be a limiting factor of the method for small institutions.

References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Kevin Duh, Yannick Estève, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutai Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Kr. Ojha, John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 2023. FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics.

Antonios Anastasopoulos, Ondřej Bojar, Jacob Bremerman, Roldano Cattoni, Maha Elbayad, Marcello Federico, Xutai Ma, Satoshi Nakamura, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Alexander Waibel, Changhan Wang, and Matthew Wiesner.

2021. FINDINGS OF THE IWSLT 2021 EVALUATION CAMPAIGN. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 1–29, Bangkok, Thailand (online). Association for Computational Linguistics.

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. 2017. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 233–242. PMLR.

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Fred Bane, Celia Soler Uguet, Wiktor Stribiżew, and Anna Zaretskaya. 2022. A comparison of data filtering methods for neural machine translation. In *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Volume 2: Users and Providers Track and Government Track)*, pages 313–325, Orlando, USA. Association for Machine Translation in the Americas.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Gabriel Bernier-Colborne and Chi-kiu Lo. 2019. NRC parallel corpus filtering system for WMT 2019. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 252–260, Florence, Italy. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Marine Carpuat, Yogarshi Vyas, and Xing Niu. 2017. Detecting cross-lingual semantic divergence for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 69–79, Vancouver. Association for Computational Linguistics.

Vishrav Chaudhary, Yuqing Tang, Francisco Guzmán, Holger Schwenk, and Philipp Koehn. 2019. Low-resource corpus filtering using multilingual sentence embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 261–266, Florence, Italy. Association for Computational Linguistics.

Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland. European Association for Machine Translation.

Miquel Esplà-Gomis, Víctor M. Sánchez-Cartagena, Jaume Zaragoza-Bernabeu, and Felipe Sánchez-Martínez. 2020. Bicleaner at WMT 2020: Universitat d'alacant-prompsit's submission to the parallel corpus filtering shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 952–958, Online. Association for Computational Linguistics.

Vitaly Feldman and Chiyuan Zhang. 2020. What neural networks memorize and why: Discovering the long tail via influence estimation. In *Advances in Neural Information Processing Systems*, volume 33, pages 2881–2891. Curran Associates, Inc.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi,

George Foster, Alon Lavie, and André F. T. Martins. 2022a. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Markus Freitag, David Vilar, David Grangier, Colin Cherry, and George Foster. 2022b. A natural diet: Towards improving naturalness of machine translation output. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3340–3353, Dublin, Ireland. Association for Computational Linguistics.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Cyril Goutte, Marine Carpuat, and George Foster. 2012. The impact of sentence alignment errors on phrase-based machine translation performance. In *Proceedings of the 10th Conference of the Association for Machine Translation in the Americas: Research Papers*, San Diego, California, USA. Association for Machine Translation in the Americas.

Christian Herold, Jan Rosendahl, Joris Vanvinckenroye, and Hermann Ney. 2022. Detecting various types of noise for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2542–2551, Dublin, Ireland. Association for Computational Linguistics.

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving translation quality by discarding most of the phrasetable. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 967–975, Prague, Czech Republic. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2018a. Dual conditional cross-entropy filtering of noisy parallel corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2018b. Microsoft's submission to the WMT2018 news translation task: How I learned to stop worrying and love the data. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 425–430, Belgium, Brussels. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 Metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019. OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122, Florence, Italy. Association for Computational Linguistics.

Huda Khayrallah and Philipp Koehn. 2018. On the impact of various types of noise on neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742, Online. Association for Computational Linguistics.

Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino. 2019. Findings of the WMT 2019 shared task on parallel corpus filtering for low-resource conditions. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 54–72, Florence, Italy. Association for Computational Linguistics.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018. Findings of the WMT 2018 shared task on parallel corpus filtering. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 726–739, Belgium, Brussels. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual

datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Chi-kiu Lo and Eric Joanis. 2020. Improving parallel data identification using iteratively refined sentence alignments and bilingual mappings of pre-trained language models. In *Proceedings of the Fifth Conference on Machine Translation*, pages 972–978, Online. Association for Computational Linguistics.

Chi-kiu Lo and Michel Simard. 2019. Fully unsupervised crosslingual semantic textual similarity metric based on BERT for identifying parallel data. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 206–215, Hong Kong, China. Association for Computational Linguistics.

Jun Lu, Xin Ge, Yangbin Shi, and Yuqi Zhang. 2020. Alibaba submission to the WMT20 parallel corpus filtering task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 979–984, Online. Association for Computational Linguistics.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Proceedings of the 5th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 135–144, Tiburon, USA. Springer.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz Rojas. 2020. Bifixer and bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal. European Association for Machine Translation.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Philip Resnik. 1999. Mining the web for bilingual text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 527–534, College Park, Maryland, USA. Association for Computational Linguistics.

Víctor M. Sánchez-Cartagena, Marta Bañón, Sergio Ortiz-Rojas, and Gema Ramírez. 2018. Prompsit's submission to WMT 2018 parallel corpus filtering shared task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 955–962, Belgium, Brussels. Association for Computational Linguistics.

Julian Schamper, Jan Rosendahl, Parnia Bahar, Yunsu Kim, Arne Nix, and Hermann Ney. 2018. The RWTH Aachen University supervised machine translation systems for WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 496–503, Belgium, Brussels. Association for Computational Linguistics.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Rico Sennrich and Martin Volk. 2011. Iterative, MT-based sentence alignment of parallel texts. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 175–182, Riga, Latvia. Northern European Association for Language Technology (NEALT).

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020. Findings of the WMT 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764, Online. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón F. Astudillo, and André F. T. Martins. 2018. Findings of the WMT 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 689–709, Belgium, Brussels. Association for Computational Linguistics.

Kaveh Taghipour, Shahram Khadivi, and Jia Xu. 2011. Parallel corpus refinement as an outlier detection algorithm. In *Proceedings of Machine Translation Summit XIII: Papers*, Xiamen, China.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 2022. UniTE: Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127, Dublin, Ireland. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Jaume Zaragoza-Bernabeu, Gema Ramírez-Sánchez, Marta Bañón, and Sergio Ortiz Rojas. 2022. Bicleaner AI: Bicleaner goes neural. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 824–831, Marseille, France. European Language Resources Association.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Appendices

A Test Data Statistics

Table 6 shows statistics of the test data sets after filtering the sentences with a length of over 128 tokens.

	test set	lines	filtered
WMT'22	en → de	2037	2036
	de → en	1984	1981
	en → ja	2037	2037
	ja → en	2008	2007
	en → zh	2037	2037
	zh → en	1875	1849
WMT'23	en → de	557	404
	de → en	549	468
	en → ja	2074	2073
	ja → en	1992	1988
	en → zh	2074	2073
	zh → en	1976	1948

Table 6: WMT test set sizes. All test sets are filtered to use less than 128 tokens. This mainly reduced the en ↔ de WMT'23 test set since this was a paragraph level task. The effect on all other test sets is minimal.

B Additional Results

B.1 WMT Data: Domain-specific Evaluation

The WMT'22 and WMT'23 test sets are comprised of text originating from different domains. The scores reported in the main text correspond to the evaluation of the corpora as a whole. In Tables 7 to 12 we show the results for each individual domain. It can be seen that the improvements are achieved over all separate domains. There are only two cases where training on all data performs slightly better than filtering with BLEURTQE (ecommerce de → en in Table 8, and manuals zh → en in Table 10).

B.2 Other Metrics

In this appendix we report the COMET22, BLEURT, BLEU and CHRF scores for the experiments reported in Section 4. Table 13 shows the results for German → English, Table 14 for English → German, Table 15 for English → Japanese, Table 16 for Japanese → English, Table 17 for English → Chinese and Table 18 for Chinese → English. The additional metrics support the conclusions of the paper.

	WMT'22				WMT'23			
	conversation	ecommerce	news	social	mastodon	news	speech	user review
None	87.9	87.4	86.4	83.1	82.5	80.8	80.9	82.0
BICLEANER	88.7	88.1	86.8	83.1	82.5	82.3	81.4	82.1
COMETKIWI	88.8	88.3	86.8	83.9	83.3	82.9	81.6	83.8
BLEURTQE	**88.9**	**88.5**	**87.2**	**84.2**	**84.2**	**83.1**	**81.8**	**84.0**

Table 7: COMET22 scores for each domain of en → de WMT test sets.

	WMT'22			
	conversation	ecommerce	news	social
None	84.7	**85.4**	84.4	83.5
BICLEANER	84.8	85.0	84.5	82.8
COMETKIWI	85.0	85.2	84.8	83.6
BLEURTQE	**85.1**	85.3	**84.9**	**83.9**

Table 8: COMET22 scores for each domain of de → en WMT test sets.

	WMT'22			
	conversation	ecommerce	news	social
None	84.2	84.0	81.5	75.3
BICLEANER	85.2	83.9	81.8	76.1
COMET22	86.0	84.8	83.3	78.1
BLEURT	**86.4**	**84.9**	**83.6**	**78.7**

Table 9: COMET22 scores for each domain of en → zh WMT test sets.

	WMT'22				WMT'23		
	conversation	ecommerce	news	social	manuals	news	user review
None	74.0	66.8	76.8	74.1	**77.7**	78.9	68.4
BICLEANER	75.2	70.8	77.9	75.6	77.4	79.5	70.5
COMETKIWI	**76.4**	71.2	78.2	75.7	77.5	**80.1**	70.6
BLEURTQE	75.9	**71.5**	**78.3**	**76.0**	77.5	79.7	**71.0**

Table 10: COMET22 scores for each domain of zh → en WMT test sets.

	WMT'22			
	conversation	ecommerce	news	social
None	88.1	87.5	86.4	80.6
BICLEANER	88.6	87.7	86.5	81.0
COMETKIWI	89.2	87.9	87.3	81.9
BLEURTQE	**89.3**	**88.6**	**87.7**	**82.5**

Table 11: COMET22 scores for each domain of en → ja WMT test sets.

	WMT'22			
	conversation	ecommerce	news	social
None	77.5	83.0	75.2	74.9
BICLEANER	77.0	83.1	77.1	75.2
COMETKIWI	77.4	**84.1**	77.8	75.4
BLEURTQE	**78.2**	83.9	**78.3**	**75.5**

Table 12: COMET22 scores for each domain of ja → en WMT test sets.

	WMT'22 (dev)				WMT'19				WMT'23			
Filter	COMET22	BLEURT	BLEU	CHRF	COMET22	BLEURT	BLEU	CHRF	COMET22	BLEURT	BLEU	CHRF
None	84.5	73.4	32.3	57.4	84.6	73.1	41.1	64.8	83.3	72.2	37.4	61.2
Random	84.2	73.1	32.0	57.1	84.3	72.9	40.5	64.4	83.5	72.2	38.2	62.1
BICLEANER	84.2	73.0	32.0	57.2	84.8	73.4	41.3	65.5	84.0	73.1	38.9	63.5
COMETKIWI	84.6	73.6	**32.5**	**57.5**	85.1	**74.0**	**41.8**	**65.7**	**84.6**	73.8	**40.6**	**65.0**
BLEURTQE	**84.8**	**73.7**	32.3	57.4	**85.2**	**74.0**	41.4	65.3	**84.6**	**73.9**	39.9	64.2

	IWSLT'21				IWSLT'23			
Filter	COMET22	BLEURT	BLEU	CHRF	COMET22	BLEURT	BLEU	CHRF
None	84.2	73.2	27.8	52.7	**87.4**	**79.1**	**47.2**	**72.4**
Random	84.1	73.0	27.3	52.5	87.2	78.5	45.6	71.2
BICLEANER	84.1	73.0	27.5	52.5	87.3	79.0	46.9	72.1
COMETKIWI	**84.4**	**73.3**	**28.0**	**53.0**	87.3	79.0	46.6	71.9
BLEURTQE	**84.4**	**73.3**	**28.0**	**53.0**	87.3	79.0	47.1	72.2

Table 13: Full results for German → English. The COMETKIWI and BLEURTQE results for IWSLT'21 are identical due to rounding.

	WMT'22 (dev)				WMT'19				WMT'23			
Filter	COMET22	BLEURT	BLEU	CHRF	COMET22	BLEURT	BLEU	CHRF	COMET22	BLEURT	BLEU	CHRF
None	86.2	76.7	35.9	62.6	86.0	75.7	43.1	67.0	81.8	70.1	38.2	61.6
Random	86.0	76.5	35.1	62.2	85.5	75.0	42.4	66.5	80.8	68.9	36.4	61.0
BICLEANER	86.7	77.5	36.4	63.1	86.4	76.2	42.9	67.2	82.2	70.5	38.4	62.3
COMETKIWI	86.9	77.6	36.2	63.0	**86.7**	76.4	**44.0**	**68.0**	82.9	71.9	**40.9**	**65.9**
BLEURTQE	**87.2**	**78.0**	**36.7**	**63.3**	**86.7**	**76.5**	42.1	66.8	**83.5**	**72.4**	**40.9**	65.5

	IWSLT'21				IWSLT'23			
Filter	COMET22	BLEURT	BLEU	CHRF	COMET22	BLEURT	BLEU	CHRF
None	83.2	73.0	23.2	56.6	84.5	76.5	45.8	72.2
Random	82.8	72.8	22.8	56.3	84.2	76.2	44.9	71.7
BICLEANER	83.3	73.1	23.1	56.7	84.9	**76.8**	45.2	71.9
COMETKIWI	83.6	73.5	23.2	56.9	84.8	76.6	**45.9**	**72.4**
BLEURTQE	**83.9**	**74.0**	**23.9**	**57.2**	**85.1**	**76.8**	45.6	72.0

Table 14: Full results for English → German.

	WMT'22 (dev)				WMT'23				IWSLT'23			
Filter	COMET22	BLEURT	BLEU	CHRF	COMET22	BLEURT	BLEU	CHRF	COMET22	BLEURT	BLEU	CHRF
None	85.6	64.5	22.2	31.4	82.3	58.3	19.0	28.9	86.9	68.1	40.5	48.1
Random	84.5	62.8	21.0	30.3	80.7	55.5	17.2	27.2	85.8	65.8	35.7	43.6
BICLEANER	86.0	64.9	21.9	31.5	83.2	59.2	19.1	29.0	87.2	68.2	39.2	47.1
COMETKIWI	86.6	65.5	**22.7**	32.1	83.7	**60.0**	19.3	**29.5**	**87.9**	**69.3**	**41.3**	**48.6**
BLEURTQE	**87.0**	**66.1**	**22.7**	**32.2**	**84.0**	**60.0**	**19.5**	**29.5**	87.4	68.1	38.8	46.5

Table 15: Full results for English → Japanese.

	WMT'22 (dev)				WMT'23				IWSLT'23			
Filter	COMET22	BLEURT	BLEU	CHRF	COMET22	BLEURT	BLEU	CHRF	COMET22	BLEURT	BLEU	CHRF
None	77.6	62.6	18.1	42.5	75.9	62.2	16.0	41.6	**85.5**	**73.4**	**30.5**	**62.2**
Random	75.9	60.8	16.1	40.8	75.0	61.0	15.0	40.7	84.6	71.9	27.4	59.9
BICLEANER	78.1	63.6	18.2	44.3	77.4	63.3	17.2	44.2	85.0	72.2	28.4	61.0
COMETKIWI	78.7	64.1	18.8	44.6	78.0	64.0	16.6	44.1	85.0	72.6	28.2	60.8
BLEURTQE	**79.0**	**64.4**	**19.0**	**45.2**	**78.2**	**64.3**	**17.4**	**44.7**	85.1	72.8	29.6	61.7

Table 16: Full results for Japanese → English.

	WMT'22 (dev)				WMT'23			
	COMET22	BLEURT	BLEU	CHRF	COMET22	BLEURT	BLEU	CHRF
None	81.2	65.8	37.7	33.5	79.7	64.8	43.4	39.0
Random	80.2	64.5	36.6	32.7	79.3	64.3	41.6	37.3
BICLEANER	81.7	66.6	37.0	33.0	80.3	65.7	42.4	37.6
COMETKIWI	83.0	68.0	38.2	34.0	**82.5**	**68.2**	**44.0**	**40.1**
BLEURTQE	**83.4**	**68.5**	**38.7**	**34.4**	82.2	67.7	43.6	38.9
	WMT'19				IWSLT'23			
	COMET22	BLEURT	BLEU	CHRF	COMET22	BLEURT	BLEU	CHRF
None	77.6	59.4	31.1	27.4	**84.2**	**72.2**	**52.5**	47.1
Random	77.2	59.1	30.7	27.5	82.1	69.3	47.6	41.9
BICLEANER	78.4	60.4	31.1	27.5	83.3	70.7	47.9	**42.3**
COMETKIWI	**80.1**	**62.2**	**32.1**	**28.3**	84.1	71.2	47.9	42.2
BLEURTQE	79.9	61.9	31.7	28.1	84.1	71.1	47.3	42.2

Table 17: Full results for English → Chinese.

	WMT'22 (dev)				WMT'23			
	COMET22	BLEURT	BLEU	CHRF	COMET22	BLEURT	BLEU	CHRF
None	72.8	59.5	17.4	46.2	74.7	61.0	18.8	44.9
Random	72.2	58.8	16.9	46.3	74.2	60.2	18.5	44.9
BICLEANER	74.8	60.9	17.1	47.6	75.7	61.5	19.1	45.8
COMETKIWI	75.2	61.4	**17.9**	**48.5**	**76.0**	61.6	**19.2**	**46.2**
BLEURTQE	**75.4**	**61.7**	17.7	48.1	**76.0**	**61.9**	18.6	45.7
	WMT'19				IWSLT'23			
	COMET22	BLEURT	BLEU	CHRF	COMET22	BLEURT	BLEU	CHRF
None	78.3	65.6	23.7	53.6	**84.9**	**74.5**	**33.3**	**63.2**
Random	78.1	65.0	23.2	53.2	84.1	73.8	31.7	62.1
BICLEANER	79.6	66.6	23.9	54.7	84.2	73.9	32.6	62.4
COMETKIWI	**80.0**	67.0	**24.7**	**55.5**	84.5	73.9	30.8	61.6
BLEURTQE	79.9	**67.2**	23.7	54.9	84.8	**74.5**	32.1	62.5

Table 18: Full results for Chinese → English.

Results of WMT23 Metrics Shared Task: Metrics might be Guilty but References are not Innocent

Markus Freitag[(1)], **Nitika Mathur**[(2)], **Chi-kiu Lo 羅致翹**[(3)], **Eleftherios Avramidis**[(4)], **Ricardo Rei**[(5,6,7)], **Brian Thompson**[(8)], **Tom Kocmi**[(9)], **Frédéric Blain**[(10)], **Daniel Deutsch**[(1)], **Craig Stewart**[(11)], **Chrysoula Zerva**[(7,12)], **Sheila Castilho**[(13)], **Alon Lavie**[(11)], **George Foster**[(1)]
[(1)]Google Research [(2)]Oracle Digital Assistant [(3)]National Research Council Canada
[(4)]German Research Center for Artificial Intelligence (DFKI) [(5)]Unbabel [(6)]INESC-ID
[(7)]Instituto Superior Técnico [(8)]AWS AI Labs [(9)]Microsoft [(10)]Tilburg University [(11)]Phrase
[(12)]Instituto de Telecomunicações [(13)]Dublin City University
wmt-metrics@googlegroups.com

Abstract

This paper presents the results of the WMT23 Metrics Shared Task. Participants submitting automatic MT evaluation metrics were asked to score the outputs of the translation systems competing in the WMT23 News Translation Task. All metrics were evaluated on how well they correlate with human ratings at the system and segment level. Similar to last year, we acquired our own human ratings based on expert-based human evaluation via Multidimensional Quality Metrics (MQM). Following last year's success, we also included a challenge set subtask, where participants had to create contrastive test suites for evaluating metrics' ability to capture and penalise specific types of translation errors. Furthermore, we improved our meta-evaluation procedure by considering fewer tasks and calculating a global score by weighted averaging across the various tasks.

We present an extensive analysis on how well metrics perform on three language pairs: Chinese→English, Hebrew→English on the sentence-level and English→German on the paragraph-level. The results strongly confirm the results reported last year, that neural-based metrics are significantly better than non-neural metrics in their levels of correlation with human judgments. Further, we investigate the impact of bad reference translations on the correlations of metrics with human judgment. We present a novel approach for generating synthetic reference translations based on the collection of MT system outputs and their corresponding MQM ratings, which has the potential to mitigate bad reference issues we observed this year for some language pairs. Finally, we also study the connections between the magnitude of metric differences and their expected significance in human evaluation, which should help the community to better understand and adopt new metrics.

Metric		avg corr
XCOMET-Ensemble	**1**	**0.825**
XCOMET-QE-Ensemble*	2	0.808
MetricX-23	2	0.808
GEMBA-MQM*	2	0.802
MetricX-23-QE*	2	0.800
mbr-metricx-qe*	3	0.788
MaTESe	3	0.782
CometKiwi*	3	0.782
COMET	3	0.779
BLEURT-20	3	0.776
KG-BERTScore*	3	0.774
sescoreX	3	0.772
cometoid22-wmt22*	4	0.772
docWMT22CometDA	4	0.768
docWMT22CometKiwiDA*	4	0.767
Calibri-COMET22	4	0.767
Calibri-COMET22-QE*	4	0.755
YiSi-1	4	0.754
MS-COMET-QE-22*	5	0.744
prismRef	5	0.744
mre-score-labse-regular	5	0.743
BERTscore	5	0.742
XLsim	6	0.719
f200spBLEU	7	0.704
MEE4	7	0.704
tokengram_F	7	0.703
embed_llama	7	0.701
BLEU	7	0.696
chrF	7	0.694
eBLEU	7	0.692
Random-sysname*	8	0.529
prismSrc*	9	0.455

Table 1: Official ranking of primary submissions to the WMT23 Metric Task. The final score is the weighted average correlation over 10 different tasks. Starred metrics are reference-free, and underlined metrics are baselines. See Table 18 for the pairwise comparisons from which the ranks were derived.

1 Introduction

The metrics shared task[1] has been a key component of WMT since 2008, serving as a way to validate the use of automatic MT evaluation metrics and drive the development of new metrics. We eval-

[1] https://wmt-metrics-task.github.io/

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 578–628
December 6–7, 2023.

uate reference-based automatic metrics that score MT output by comparing the translations with a reference translation generated by human translators, who are instructed to translate "from scratch" without post-editing from MT. In addition, we also invited submissions of reference-free metrics (quality estimation metrics or QE metrics) that compare MT outputs directly with the source segments. All metrics are evaluated based on their agreement with human rating when scoring MT systems and human translations at the system and sentence level. The final ranking of this year's submitted primary metrics is shown in Table 1. Below are some key details and changes for this year's metric shared task:

- **Language Pairs**: For this year, we focus on three main language pairs: (i) One language pair with paragraph-level test sets: English→German (en→de), (i) one low-resource language pair with sentence-level test sets: Hebrew→English (he→en), (iii) one high-resource language pair with sentence-level test sets: Chinese→English (zh→en).

- **Human Evaluation**: Like last year, we collected our own human ratings for our three language pairs from professional translators via MQM (Lommel et al., 2014; Freitag et al., 2021). We released and uploaded[2] all MQM annotations, and we recommend using Marot[3] for looking into this data.

- **Meta Evaluation**: This year's meta-evaluation is significantly streamlined from last year's. Instead of 201 tasks, we use just 10, designed to capture complementary ranking and linearity properties at system- and segment-level granularity. We replace Kendall's tau at the segment level with a version of pairwise accuracy that gives metrics credit for correctly predicting ties in human scores, while automatically calibrating for each metric's natural scale (Deutsch et al., 2023). Instead of averaging per-task ranks to derive an overall score for each metric, we simply average correlation/accuracy scores across tasks. This places metric scores on an absolute scale, and makes them independent of the performance of other metrics. Finally, we compute top-level significance clusters to provide a clearer global ranking of participating metrics.

- **Synthetic Reference**: The MQM scores for the human reference translation for zh→en were unexpectedly low, ranking humans below almost all WMT submissions. We investigate the impact of bad reference translations on reference-based metrics and propose a novel approach to create a synthetic reference translation from all WMT submissions and their corresponding MQM scores.

- **Challenge Sets Subtask**: For the second year, we include a decentralized sub-task on challenge sets, in which test sets are submitted by different research teams targeting to reveal metrics' abilities or the weaknesses in evaluating particular translation phenomena. We received three challenge sets covering a wide range of translation errors and linguistic phenomena in more than a hundred translation directions.

- **Understand Magnitude of Score Difference**: This year, we include two analyses to understand the meaning of the score differences that metrics present with respect to the statistical significance of MT system rankings according to human annotations and metric scores. These analyses provide additional assistance for MT researchers to build an intuition on the relationship between the magnitude of metric score differences and the reliability of the improved translation quality.

- **MTME**: Similar to last year, all the data has been uploaded to MTME[4], and all results in this paper are calculated with this analysis tool. We encourage every metric developer to use MTME to calculate contrastive scores to enhance consistency and comparability going forward.

Our main findings are:

- XCOMET-Ensemble is the winner of the WMT23 Metrics Shared Task (Table 1).

- High correlations between automatic metrics and human judgments at the segment level do not necessarily guarantee high correlations at the system level (Figure 5).

[2] https://github.com/google/wmt-mqm-human-evaluation

[3] https://github.com/google-research/google-research/tree/master/marot

[4] https://github.com/google-research/mt-metrics-eval

- Reference quality matters: The low quality reference for zh→en significantly lowered the correlation of all metrics with human judgement (Section 8).
- We determined the magnitude of score differences required to produce a statistically significant difference in human judgment for each metric, revealing that even minor score differences of the top performing metrics can be statistical significant with high probability (Section 7).
- Results from the challenge sets independently agreed with our findings that the quality of reference matters. Developing reference-free metrics is worth further exploration, and metric researchers are advised to investigate into the influence of language-agnostic multilingual embeddings on MT evaluation. It is equally important for metric researchers to test the performance of metrics in diverse collection of linguistic phenomena and wider landscape of translation quality in order to minimize unexpected behaviours of metrics (Section 10).

The rest of the paper is organized as follows: Section 2 describes the test data and additional MT systems that we trained. Section 3 presents an overview of the conducted expert-based human evaluation. Section 4 describes the metrics evaluated this year (baselines and participants). Section 5 describes the conducted meta-evaluation. Section 6 reports our main results. Section 7 interprets and evaluates metrics' scores beyond correlations. Section 8 analyses the impact of bad reference translations on the various metrics. Section 9 summarizes our results for additional WMT23 Translation task language-pairs based on their Direct Assessment human evaluation. Section 10 presents a description of the submitted challenge sets along with their findings. Finally, Section 11 presents our most relevant conclusions.

2 Translation Systems

Similar to previous years' editions, the source, reference texts, and MT system outputs for the metrics task are mainly derived from the WMT23 General MT Shared Task. In addition to the MT system outputs from the WMT evaluation campaign, we included translations from two additional MT systems which we deemed interesting for evaluation.

2.1 WMT Test Sets

We use test sets prepared by the WMT23 General MT Shared Task (Kocmi et al., 2023). For our three main language pairs, the test sets contain 557 en→de, 1910 he→en, and 1976 zh→en segments. This year, the test sets cover up to five domains from the following list: news, conversational, user reviews, manuals, and social. Each language pair contains a comparable number of sentences from each domain, resulting in reasonably balanced test sets.

English→German contains four balanced domains: news, social, conversational, and user reviews. In contrast to other language pairs, segments are paragraphs rather than sentences.

Hebrew→English contains only news and user reviews domains. This language pair has two human references, but one of them (refA) is suspected of being a post-edited Online-B system output.

Chinese→English contains news, user reviews, and manuals. The first two domains contain around 750 sentences, while manuals contains around 500.

The reference translations provided for the test sets are produced by professional translators.

For more details regarding the news test sets, we refer the reader to the WMT23 General MT Shared Task findings paper (Kocmi et al., 2023).

2.2 Additional MT Output

Similar to last year, we made an effort to expand the pool of translations beyond the WMT submissions, which can potentially be quite similar to each other. We added translations which we expected to differ in two main ways from the submissions: 1) by using a massively multilingual model; and 2) by generating with MBR decoding;

For our multilingual model, we selected the 3.3B parameter NLLB200 model (NLLB Team et al., 2022) via the huggingface (Wolf et al., 2020) interface. We found NLLB200 to significantly outperform the M2M100 (Fan et al., 2021) that we used last year.

Minimum Bayes Risk (MBR) decoding has recently gained attention in MT as a decision rule, with the potential to overcome some of the biases of MAP decoding in NMT (Eikema and Aziz, 2020; Müller and Sennrich, 2021; Eikema and Aziz, 2021; Freitag et al., 2022; Fernandes et al., 2022). MBR decoding centrally relies on a reference-based utility metric: its goal is to identify a hypothesis with a high estimated utility (expectation

under model distribution) with the hope that a high estimated utility translates into a high actual utility (with respect to a human reference). In practice, this means generating several candidate translations and finding the translation that is most similar to the rest of the candidate translations.

We produced both the top-1 greedy translation and MBR outputs. For MBR, we sampled 100 translation candidates from the model via Epsilon sampling (Hewitt et al., 2022; Freitag et al., 2023). We used `epsilon_cutoff=0.02` and `eta_cutoff=0.0`. This year, we used sentence-level BLEU from sacreBLEU (Post, 2018) with the default 'a13' tokenizer and the 'floor' smoothing method as utility function only.

3 MQM Human Evaluation

Automatic metrics are usually evaluated by measuring correlations with human ratings. The quality of the underlying human ratings is critical, and recent findings (Freitag et al., 2021) have shown that crowdsourced human ratings are not reliable for high quality MT output. Furthermore, an evaluation schema based on MQM (Lommel et al., 2014), which requires explicit error annotation, is preferable to an evaluation schema that only asks raters for a single scalar value per translation. Similar to last year, we decided to conduct our own MQM-based human evaluation on a subset of submissions and language pairs that are most interesting for evaluating current metrics.

MQM is a general framework that provides a hierarchy of translation errors which can be tailored to specific applications. Google and Unbabel sponsored the human evaluation for this year's metrics task for a subset of language pairs using either professional translators (English→German, Chinese→English) or trusted and trained raters (Hebrew→English). The error annotation typology and guidelines used by Google's and Unbabel's annotators differ slightly and are described in the following two sections.

3.1 English→German and Chinese→English

Annotations for English→German and Chinese→English were sponsored and executed by Google, using 18 professional translators (10 for English→German, 8 for Chinese→English) having access to the full document context. Each segment gets annotated by a single rater. Instead of assigning a scalar value to each translation, annotators were instructed to label error spans within each segment in a document, paying particular attention to document context. Each error was highlighted in the text, and labelled with an error category and a severity. Segments that are too badly garbled to permit reliable identification of individual errors are assigned a special *Non-translation* error. Error severities are assigned independent of category, and consist of *Major*, *Minor*, and *Neutral* levels, corresponding respectively to actual translation or grammatical errors, smaller imperfections and purely subjective opinions about the translation. Since we are ultimately interested in scoring segments, we adopt the weighting scheme shown in Table 2. For more details, exact annotator instructions and a list of error categories, we refer the reader to Freitag et al. (2021) as the exact same setup was used for the previous two metrics tasks.

Severity	Category	Weight
Major	Non-translation	25
	all others	5
Minor	Fluency/Punctuation	0.1
	all others	1
Neutral	all	0

Table 2: Google's MQM error weighting.

3.2 Hebrew→English

The annotations for the Hebrew→English language pair were sourced from Unbabel, who engaged four professional native language annotators possessing extensive translation experience. Much like Google's approach, these annotators were provided with the full document context, comprising up to ten segments. Their task was to identify and classify errors by highlighting them, following Unbabel's MQM 3.0 typology[5].

The annotators were instructed to classify the errors based on severity, with Unbabel's classification encompassing not only "Minor" and "Major" error severities (analogous to Google's criteria) but also a "Critical" error severity. However, to ensure consistency in our evaluation process, we opted to align with the Google methodology outlined previously. Specifically, we treated all annotated "Critical" errors as "Major" errors, and we applied a weighting scheme for punctuation errors, as detailed in Table 2.

[5]see Unbabel Annotation Guidelines - Typology 3.0

3.3 Human Evaluation Results

Due to the fact that we ran our own human evaluation, we were only able to evaluate a subset of the test segments. In Table 3, you can see the number of segments and documents for each language pair and test set that we used for human evaluation. We followed a simple and consistent approach to downsample the data: we considered each document, while only keeping the first 10 sentences of each document. By doing this, we did not need to discard most of the documents and only needed to crop longer documents. The English→German test is on the paragraph-level, and we had to discard two documents as the first paragraph already contained more than 10 sentences. In all cases, the MQM score for a segment is the sum of the scores for the errors in that segment, and the MQM score for a test set is the average of the MQM scores of the segments that were annotated.

The results of the MQM human evaluation can be seen in Table 4. Most of the reference translations are ranked first, except for refA for Chinese→English. Not ranking the human evaluation on top of the MT output is usually a signal for a corrupt human evaluation. We double-checked the annotation for refA and can confirm that the reference translation indeed contained many errors.

4 Baselines and Submissions

We computed scores for several baseline metrics in order to compare submissions against previous well-studied metrics. We will start by describing those baselines, and then we will describe the submissions from participating teams. An overview of the evaluated metrics can be seen in Table 5.

4.1 Baselines

SacreBLEU baselines We use the following metrics from SacreBLEU (Post, 2018) as baselines:

- **BLEU (Papineni et al., 2002)** is based on the precision of n-grams between the MT output and its reference weighted by a brevity penalty. Using SacreBLEU we obtained sentenceBLEU values using the `sentence_bleu` Python function and for corpus-level BLEU we used `corpus_bleu` (both with default arguments[6]).

- **F200SPBLEU (NLLB Team et al., 2022)** are BLEU scores computed with subword tokenization done by the standardized FLORES-200 Sentencepiece models. We used the command line SacreBLEU to compute the sentence level F200SPBLEU[7] and we average the segment-level scores to obtain a corpus-level score.

- **CHRF (Popović, 2015)** uses character n-grams instead of word n-grams to compare the MT output with the reference. For CHRF we used the SacreBLEU `sentence_chrf` function (with default arguments[8]) for segment-level scores and we average those scores to obtain a corpus-level score.

BERTSCORE (Zhang et al., 2020) leverages contextual embeddings from pre-trained transformers to create soft-alignments between words in candidate and reference sentences using cosine similarity. Based on the alignment matrix, BERTSCORE returns a precision, recall and F1 score. We used F1 without TF-IDF weighting.

BLEURT (Sellam et al., 2020) is a learned metric fine-tuned on Direct Assessments (DA). Unlike COMET, BLEURT encodes the translation and the reference together and utilizes the `[CLS]` token as an embedding to represent the pair. We employed the BLEURT20 checkpoint (Pu et al., 2021), which was trained on top of RemBERT using DA data from previous shared tasks spanning from 2015 to 2019, along with additional synthetic data created from Wikipedia articles.

COMET (Rei et al., 2022a) is a learned metric fine-tuned using DA from previous WMT Translation shared tasks. This metric relies on sentence embeddings from the source, translation, and reference to produce a final score. We utilized the default model `wmt22-comet-da` provided in version 2.0.2 of the `Unbabel/COMET` framework. This model employs XLM-R large as its backbone model and is trained on data from the 2017 to 2019 WMT shared tasks, in combination with the MLQE-PE corpus (Fomicheva et al., 2022).

COMETKIWI (Rei et al., 2022b) is a reference-free learned metric that functions similarly to

[6] lnrefs.1|case.mixed|lang.LANGPAIR|tok.13a|smooth.exp| version.2.3.0

[7] nrefs:1|case:mixed|eff:yes|tok:flores200|smooth:exp| version:2.3.0

[8] chrF2|lang.LANGPAIR|nchars.6|space.false|version.2.3.0

language	news	social	speech	user reviews	manuals
en→de	104/139 (30/30)	206/212 (79/79)	58/113 (23/25)	92/93 (58/58)	n/a
he→en	619/1558 (68/70)	n/a	n/a	201/352 (26/26)	n/a
zh→en	377/763 (38/38)	n/a	n/a	677/726 (127/127)	123/487 (14/14)

Table 3: Numbers of MQM-annotated segments per domain (number of docs in brackets).

English→German ↓ System	all	news	social	speech	user-reviews
refA	2.96	3.12	2.02	4.74	3.77
GPT4-5shot	3.72	4.00	2.41	6.51	4.60
ONLINE-W	3.95	2.69	2.62	5.90	7.13
ONLINE-B	4.71	4.35	3.14	5.96	7.85
ONLINE-Y	5.64	4.45	3.67	7.48	10.26
ONLINE-A	5.67	4.40	3.84	7.78	9.87
ONLINE-G	6.57	6.43	4.12	7.93	11.38
ONLINE-M	6.94	4.87	4.41	8.30	14.08
Lan-BridgeMT	8.67	7.99	5.55	9.72	15.78
LanguageX	9.25	8.43	5.74	14.23	14.92
NLLB_Greedy	9.54	8.29	5.20	14.82	17.35
NLLB_MBR_BLEU	10.79	9.93	5.53	17.75	19.18
AIRC	14.23	14.32	8.34	20.34	23.45

Hebrew→English ↓ System	all	news	user-reviews
refA	1.17	1.28	0.86
GPT4-5shot	1.33	1.29	1.48
ONLINE-A	1.38	1.34	1.50
ONLINE-B	1.55	1.60	1.39
GTCOM_DLUT	1.89	1.85	1.99
UvA-LTL	1.92	1.80	2.30
ONLINE-G	2.06	2.06	2.04
ONLINE-Y	2.35	2.42	2.12
LanguageX	2.38	2.33	2.53
Samsung_Research_Philippines	3.23	3.62	2.05
NLLB_MBR_BLEU	3.68	3.83	3.20
NLLB_Greedy	3.79	3.98	3.19
Lan-BridgeMT	3.79	3.81	3.74

Chinese→English ↓ System	all	news	manuals	user-reviews
Lan-BridgeMT	2.10	2.31	1.28	2.13
GPT4-5shot	2.31	2.26	2.01	2.39
Yishu	3.23	3.34	1.67	3.46
ONLINE-B	3.39	3.27	1.78	3.74
HW-TSC	3.40	3.40	1.83	3.68
ONLINE-A	3.79	2.90	1.83	4.63
ONLINE-Y	3.79	3.47	2.84	4.14
ONLINE-G	3.86	3.58	2.02	4.34
ONLINE-W	4.06	3.84	2.16	4.53
LanguageX	4.23	4.05	2.84	4.59
IOL_Research	4.59	3.60	1.85	5.63
refA	4.83	5.04	5.17	4.65
ONLINE-M	5.43	4.71	2.98	6.28
ANVITA	6.08	5.17	2.97	7.15
NLLB_MBR_BLEU	6.36	6.57	3.39	6.78
NLLB_Greedy	6.57	6.70	2.95	7.16

Table 4: MQM human evaluations for generalMT2023. Lower average error counts represent higher MT quality. Systems above any solid line are significantly better than those below, based on all domains with $p < 0.05$.

BLEURT, but instead of encoding the translation along with its reference, it uses the source. We utilized the `wmt22-cometkiwi-da` model, which was a top-performing reference-free metric from last year's shared task. This reference-free metric is fine-tuned on the same data as `wmt22-comet-da` using the version 2.0.2 of the `Unbabel/COMET` framework.

DOCWMT22COMETDA (Vernikos et al., 2022) is the document-level version of `wmt22-comet-da`, which computes the BERT embeddings using multi-sentence context instead of just the single sentence.

DOCWMT22COMETKIWIDA is the document-level version of WMT22-COMETKIWI-DA (QE) which computes the BERT embeddings using multi-sentence context instead of just the single sentence.

MS-COMET-QE-22 (Kocmi et al., 2022b) is built on top of COMET by Microsoft Research using proprietary data. This metric is trained on a several times larger set of human judgements compared to COMET-baseline, covering 113 languages and 15 domains. Furthermore, the authors propose filtering of human judgement with potentially low quality to further improve the model. The metric calculated scores in quality estimation fashion with only source segment and MT hypothesis.

PRISMREF and PRISMSRC (Thompson and Post, 2020a,b) PRISMREF is the reference-based PRISM that uses a multilingual MT model in zero-shot paraphrase model to score the candidate translation conditioned on the reference sentence, and the reference sentence conditioned on the candidate translation, and averages the two scores. PRISM-SRC is the source-based (i.e. QE as a metric) PRISM that uses a multilingual MT model to force-decode and score the candidate translation conditioned on the source sentence.

RANDOM-SYSNAME is a random metric that takes the system name as the only parameter. For each translation system, the metric computes the mean value X as $sha256(sysname)[0]\%10$. It uses discrete scores. Segment-level scores follow

	metric	broad category	supervised	ref. free	citation	availability (https://github.com/)
baselines	BLEU	lexical overlap			Papineni et al. (2002)	mjpost/sacrebleu
	F200SPBLEU	lexical overlap			NLLB Team et al. (2022)	mjpost/sacrebleu
	CHRF	lexical overlap			Popović (2015)	mjpost/sacrebleu
	BERTSCORE	embedding similarity			Zhang et al. (2020)	Tiiiger/bert_score
	BLEURT	fine-tuned metric	✓		Sellam et al. (2020)	google-research/bleurt
	COMET	fine-tuned metric	✓		Rei et al. (2022a)	Unbabel/COMET
	COMETKIWI	fine-tuned metric	✓	✓	Rei et al. (2022b)	Unbabel/COMET
	DOCWMT22COMETDA	fine-tuned metric	✓		Vernikos et al. (2022)	amazon-research/doc-mt-metrics
	DOCWMT22COMETKIWIDA	fine-tuned metric	✓	✓	Vernikos et al. (2022)	amazon-research/doc-mt-metrics
	MS-COMET-QE-22	fine-tuned metric	✓	✓	Kocmi et al. (2022b)	MicrosoftTranslator/MS-Comet
	PRISMREF	MT-model metric	✓		Thompson and Post (2020a,b)	thompsonb/prism
	PRISMSRC	MT-model metric	✓	✓	Thompson and Post (2020a,b)	thompsonb/prism
	RANDOM-SYSNAME	random baseline		✓	—	(not available)
	YISI-1	embedding similarity			Lo (2019)	chikiulo/yisi
primary submissions	CALIBRI-COMET22	fine-tuned metric	✓		—	(not available)
	CALIBRI-COMET22-QE	fine-tuned metric	✓	✓	—	(not available)
	COMETOID22-WMT22	fine-tuned metric	✓	✓	Gowda et al. (2023)	marian-nmt/marian-dev
	EBLEU	embedding similarity			ElNokrashy and Kocmi (2023)	munael/ebleu-mt-metrics-wmt23
	EMBED_LLAMA	embedding similarity			Dreano et al. (2023a)	SorenDreano/embed_llama
	GEMBA-MQM	LLM prompt-based metric		✓	Kocmi and Federmann (2023)	MicrosoftTranslator/GEMBA
	KG-BERTSCORE	embedding similarity		✓	Wu et al. (2023)	(not available)
	MATESE	fine-tuned metric	✓		Perrella et al. (2022)	SapienzaNLP/MaTESe
	MBR-METRICX-QE	fine-tuned metric	✓	✓	Naskar et al. (2023)	(not available)
	MEE4	lexical & embedding similarity			Mukherjee and Shrivastava (2023)	AnanyaCoder/WMT22Submission
	METRICX-23	fine-tuned metric	✓		Juraska et al. (2023)	google-research/metricx
	METRICX-23-QE	fine-tuned metric	✓	✓	Juraska et al. (2023)	google-research/metricx
	MRE-SCORE-LABSE-REGULAR	fine-tuned metric	✓		Viskov et al. (2023)	NL2G/efficient-llm-metrics
	SESCORE	fine-tuned metric	✓		Xu et al. (2022)	xu1998hz/SEScore
	TOKENGRAM_F	lexical overlap			Dreano et al. (2023b)	SorenDreano/tokengram_F
	XCOMET-ENSEMBLE	fine-tuned metric	✓		Guerreiro et al. (2023)	Unbabel/COMET
	XCOMET-QE-ENSEMBLE	fine-tuned metric	✓	✓	Guerreiro et al. (2023)	Unbabel/COMET
	XLSIM	fine-tuned metric	✓		Mukherjee and Shrivastava (2023)	AnanyaCoder/XLsim

Table 5: Baseline metrics and primary submissions for the metrics task. We categorize metrics into 3 major classes: lexical, embedding similarity and fine-tuned metrics. Regarding fine-tuned metrics we have metrics that use human quality scores such as DA or MQM and metrics that use synthetic labels for fine-tuning (3rd column).

Gaussian distribution around mean value X (in a range 0-9) and standard deviation of 2.

YiSi-1 (Lo, 2019) is a MT evaluation metric that measures the semantic similarity between a machine translation and human references by aggregating the IDF-weighted lexical semantic similarities based on the contextual embeddings extracted from pre-trained language models (e.g. RoBERTa, CamemBERT, XLM-RoBERTa, etc.).

4.2 Metric Submissions

The rest of this section summarizes the participating metrics.

CALIBRI-COMET22 and CALIBRI-COMET22-QE apply a post-processing approach to ratings provided by COMET. It uses `Unbabel/wmt22-comet-da` as the backbone for the referenced CALIBRI-COMET22 and `Unbabel/wmt22-cometkiwi-da` as the backbone for the unreferenced CALIBRI-COMET22-QE metric. The information whether a translation is error-free from MQM ratings (e.g. under Google's MQM error weighting, error-free translations have a score of 0) can be recovered. It then aims to calibrate the scores of the backbone model with respect to this binary error-freeness label using isotonic regression. During test time, it takes the samples for a given tuple (lang-pair, test-set, domain, ref, system-id) and employs a heuristic strategy to select samples from previous years that match the test sample score distribution. It then fits an isotonic regression model to the selected samples and transforms the test scores accordingly. The main idea is that in this way, the averaged system-level score can be interpreted as the fraction of error-free translations.

COMETOID22 (Gowda et al., 2023) is a reference-free metric created using knowledge distillation from reference-based metrics. Using COMET-22 as a teacher metric, it scores the MT outputs submitted to the WMT News/General Machine Translation task since 2009. A student metric, called COMETOID22, is then trained to mimic the teacher scores without using reference translation. The student metric has the same architecture as COMET-QE, and is initialized with pretrained weights from InfoXLM, a multilingual language model. We submit three variants: COMETOID22-WMT{21,22,23}, where the suffix indicates the training data cut-off year.

COMETKIWI XL/XXL (Rei et al., 2023) shares the same architecture as the COMETKIWI baseline but replaces InfoXLM with XLM-R XL (3.5B) and XXL (10.7B). In terms of training data, these models are trained on the same dataset as COMETKIWI, along with newly released Direct Assessments (DA) for Indian languages, which were introduced as additional training data for this year's Quality Estimation (QE) shared task (Blain et al., 2023).

EBLEU (ElNokrashy and Kocmi, 2023) String-based metrics such as BLEU and CHRF depend on string similarity as proxy for meaning similarity between candidate and target sentences. EBLEU stands for 'Embedded BLEU' and is loosely inspired by it. In EBLEU, we match candidate and target tokens approximately using non-contextual word embeddings and a word-to-word similarity map in a form we have dubbed "relative meaning diffusion tensors".

EMBED_LLAMA (Dreano et al., 2023a) relies on pretrained Llama2 embeddings, without any fine-tuning, to transform sentences into a vector space that establishes connections between geometric and semantic proximities. This metrics draws inspiration from Word2vec, and utilizes cosine distance for the purpose of estimating similarity or dissimilarity between sentences.

GEMBA-MQM (Kocmi and Federmann, 2023) is a LLM-enabled metric for error quality span marking. It uses three-shot prompting with the GPT4 model. In contrast to EAPrompt (Lu et al., 2023), it does not require language specific examples and requires only a single prompt.

HWTSC-EE-METRIC and KG-BERTSCORE (Wu et al., 2023) EE stands for Entropy Enhanced MT Metrics and aims at achieving a more balanced system-level rating by assigning weights to segment-level scores produced by MT metrics. The weights are determined by the difficulty of a segment determined by the entropy between the hypothesis-reference pair. This year, the COMET metric is utilized as the backbone of our EE metrics. The model we use is WMT22-COMET-DA.

KG-BERTSCORE incorporates multilingual knowledge graph into BERTSCORE and generates the final evaluation score by linearly combining the results of KGSCORE and BERTSCORE, in which

we use COMET-QE to calculate BERTSCORE this year.

MATESE (Perrella et al., 2022) leverages transformer-based encoders to identify error spans in translations, and classify their severity between Minor and Major. Differently from last year's version, MATESE is now based on DeBERTa for evaluating translations towards English, and InfoXLM for German and Russian. Furthermore, it has been re-trained using also the MQM data released at WMT2022.

MBR-METRICX-QE (Naskar et al., 2023) MBR decoding with neural utility metrics like BLEURT is known to be effective in generating high quality machine translations. We use the underlying technique of MBR decoding and develop an MBR based reference-free quality estimation metric. Our method uses an evaluator machine translation system and a reference-based utility metric (specifically BLEURT and METRICX) to calculate a quality estimation score of a model. We report results related to comparing different MBR configurations and utility metrics.

MEE4 (Mukherjee and Shrivastava, 2023) is an unsupervised, reference-based metric (an improved version of MEE) focusing on computing contextual and syntactic equivalences, along with lexical, morphological, and semantic similarity. The goal is to comprehensively evaluate the fluency and adequacy of MT outputs while also considering the surrounding context. Fluency is determined by analysing syntactic correlations, while context is evaluated by comparing sentence similarities using sentence embeddings. The ultimate score is derived from a weighted amalgamation of three distinct similarity measures: a) Syntactic similarity, which is established using a modified BLEU score. b) Lexical, morphological, and semantic similarity, quantified through explicit unigram matching. c) Contextual similarity, gauged by sentence similarity scores obtained from the Language-Agnostic BERT model.

METRICX-23 and METRICX-23-QE (Juraska et al., 2023) are learned reference-based and reference-free (respectively) regression metrics based on the mT5 encoder-decoder language model. They further fine-tune the mT5-XXL checkpoint on direct assessment data from 2015-2020 and MQM data from 2020 to 2021 as well as synthetic data. There are two contrastive submissions, "b" and "c", for both the reference-based and QE metrics. The "b" variant additionally trains on MQM data from 2022 and the "c" variant uses the PaLM-2 language model (Anil et al., 2023) to initialize the metric instead of mT5.

MRE-SCORE (Viskov et al., 2023) is a trained metric that is based on the encoder part of mT0-large model. We use a concatenation of source, reference and hypothesis texts for input. Additionally, some of the variants of the model uses contextual embeddings from LaBSE.

SESCOREX (Xu et al., 2023b) and INSTRUCTSCORE (Xu et al., 2023c) SESCOREX is an improved version of SESCORE2 (Xu et al., 2023a). Building upon the established strengths of SESCORE2, we utilize its framework for synthetic data generation to pre-train our scoring model. To further elevate the performance of SESCOREX, we introduce two key modifications: fine-tuning human rating data and transitioning the scoring backbone model to the MT5-xl model. INSTRUCTSCORE is an open-source, explainable evaluation metric for text generation. Utilizing explicit human guidelines and GPT4's implicit knowledge, we fine-tune an Llama model to provide evaluation metrics along with diagnostic reports that align with human assessments. Unlike traditional neural metrics, INSTRUCTSCORE evaluates text generation by providing a quality score based on detailed error explanations.

SLIDE (Raunak et al., 2023) Building metrics explicitly for document-level MT quality estimation has been challenging owing to the lack of large-scale document-level human annotated datasets. In this submission, we present a metric named SLIDE (Sliding Document Evaluator), which operates at the span of multiple sentences or paragraphs by way of an overlapping sliding window. SLIDE feeds each chunk into a source-based COMET model, with scores over overlapping chunks accumulated to produce a system-level score. SLIDE is motivated by two ideas: (1) Since COMET's underlying encoder is trained on wider contexts, we might observe generalizable evaluation behaviour beyond typical sentences-level lengths, within certain length limits and (2) since a sentence's evaluation will differ at different positions within a document, it may be helpful to evaluate each sentence in multiple different contexts.

TOKENGRAM_F (Dreano et al., 2023b) is an F-score-based evaluation metric for machine translation that is heavily inspired by CHRF++. By replacing word-grams with token-grams obtained from contemporary tokenization algorithms, TOKENGRAM_F captures similarities between words sharing the same semantic roots and thus obtains more accurate ratings.

XCOMET-XL/XXL (Guerreiro et al., 2023) is a new COMET (Rei et al., 2020) model that is designed to identify error spans in sentences and generate a final quality score, making it a more interpretable learnt metric. This metric is optimized for both regression and sequence tagging, and it can be used with or without references. XCOMET-QE submission results from the same model but running inference without a reference. These models utilize XLM-R XL or XXL as their backbone models, with XCOMET-XL having 3.5B parameters and XCOMET-XXL having 10.7B parameters. The training process for this metric occurs in stages, starting with DAs and then is fine-tuning on MQM data. XCOMET-ENSEMBLE is an ensemble between 1 XL and 2 XXL checkpoints that result from the different training stages.

XLSIM (Mukherjee and Shrivastava, 2023) is a supervised reference-based metric that regresses on human scores provided by WMT (2017-2022). Using a cross-lingual language model XLM-RoBERTa, we train a supervised model using a Siamese network architecture with cosine similarity loss.

5 Meta Evaluation

Our main goal in evaluating metrics is to establish a ranking that reflects a metric's performance across a range of settings and applications. Combining results from different settings is challenging because correlations with human gold scores have different ranges and may be subject to differing degrees of noise. There are also many ways of measuring correlation, with different strengths and weaknesses, and it is often not clear which is best in a given setting.

Last year, our approach was to define a large number of "tasks" (201 in total) that varied along dimensions such as language pair, domain, granularity, correlation statistic, etc. For each task, we used pairwise significance tests to establish a dense clustered ranking of participating metrics (e.g., 1, 1, 1, 2, 3, 3, ...). Motivated by theoretical results pertaining to combining rankings from different knowledge sources (Colombo et al., 2022; Dwork et al., 2001), we established an overall ranking by simply averaging the per-task ranks.

This approach has several disadvantages. First, it is difficult to incorporate new metrics into the comparison, since this requires not only computing the score of a new metric on 201 tasks, but also comparing it to all existing metrics on each task using expensive resampling significance tests. Adding a new metric also has the undesirable effect of potentially causing other metrics to swap places in the overall rankings. While rank averaging has theoretical underpinnings, as noted above, these apply to settings in which the constituent tasks provide only ranking information themselves. In order to take advantage of richer information available from correlation statistics, we derived dense ranks from pairwise significance tests, but this relies on an ad hoc clustering algorithm, and it is not clear to what extent our average ranks are supported by the original theory. They also lack confidence information, making it difficult to quantify conclusions about the overall superiority of one metric over another.

This year we adopted a much simpler approach in order to address these difficulties. We use just 10 main tasks, and compute an overall score by taking a weighted average of results from each task. We perform significance tests on each pair of metrics for each task as before, but also do so for each pair of metrics on the overall average score, allowing us to establish a clearer global ranking. The average score for a new metric can be computed relatively quickly, and it does not affect the scores of other metrics. Significance tests still require the expensive step of comparing to all other metrics, but they are no longer necessary for computing a metric's raw overall score.

We acknowledge that this approach is not perfect. One problem is that we need to combine correlations and accuracies that may have different dynamic ranges. For example, the mean Pearson correlation across all metrics for en→de at the system level is 0.88 with standard deviation 0.24, while at the segment level it is 0.39 with a standard deviation of 0.17. Averaging system-level and segment-level correlations will therefore effectively upweight the system-level contribution. We experimented with different weightings to compensate for this, but found that they did not make a large differ-

language	ref used	scored ref
en→de	A	–
he→en	B	A
zh→en	A	–

Table 6: Use of reference translations.

ence, and decided to use equal weights for simplicity. Another problem is that we do not account for dependencies among tasks. Although all tasks are at least somewhat complementary, many–such as system-level and segment-level correlations—are based on the same underlying data, and thus violate the assumptions of our hypothesis tests. We leave more sophisticated inference approaches such as proposed by Dror et al. (2017) or Hagmann and Riezler (2023) for future work.

5.1 Task Attributes

Tasks are identified by unique value assignments for each of the following attributes: language, level, and correlation statistic. Unlike last year, we no longer have tasks specific to different domains, as domains differ across languages this year. We also drop the "include-human" vs "no-human" distinction, and always score reference translations that are not used by the metrics. As shown in Table 6, Hebrew→English is the only language pair for which such a reference is available. Finally, last year we used three different averaging methods for each correlation statistic at the segment level; this year we choose only one method for each segment-level correlation.

Attributes are as follows:

Language

Language pairs include those for which we have MQM ratings—English→German, Hebrew→English, and Chinese→English—plus *all*, which indicates all pairs pooled together.

Level

We computed correlations at the *system* level and the *segment* level. For English→German, segments are paragraphs; for the two other language pairs, they are sentences. System-level scores for human ratings and for all metrics that did not supply an explicit system-level score are averages over segment-level scores.

Correlation/accuracy

We computed three correlation/accuracy statistics selected to provide complementary information:

task	lang	level	correlation	wt
1	all	system	accuracy	3
2	en→de	system	Pearson	1
3	en→de	segment	Pearson	1
4	en→de	segment	acc^*_{eq}	1
5	he→en	system	Pearson	1
6	he→en	segment	Pearson	1
7	he→en	segment	acc^*_{eq}	1
8	zh→en	system	Pearson	1
9	zh→en	segment	Pearson	1
10	zh→en	segment	acc^*_{eq}	1

Table 7: Tasks and weighting.

- System-level pairwise ranking *accuracy* (as proposed by Kocmi et al., 2021). This is computed over data pooled across all three language pairs.
- Segment-level pairwise ranking accuracy with tie calibration (as proposed by Deutsch et al., 2023). We use the acc^*_{eq} variant to compare vectors of metric and gold scores for each segment, then average the results over segments.
- System- and segment-level Pearson correlation. At the segment level, we flatten matrices of system × segment scores into vectors before comparing them.

5.2 Tasks and Weighting

Table 7 shows the complete list of tasks and their weights. All tasks receive a weight of 1, except for system-level accuracy, which has a weight of 3 because it combines data from all three language pairs.

To compute a global score for each metric across all tasks, we first map Pearson correlations from their natural range of $[-1, 1]$ into the $[0, 1]$ range of the accuracy scores, then take a weighted average of the results.

5.3 Rank Assignment

For each task, we assign ranks to metrics based on their significance clusters. To do so, we compare all pairs of metrics and determine whether the difference in their correlation scores is significant, according to the PERM-BOTH hypothesis test of Deutsch et al. (2021). We use 1000 re-sampling runs and set $p = 0.05$. As advocated by Wei et al.

(2022), we divide the sample into blocks of 100, compute significance after each block (cumulative over all blocks sampled so far), and stop early if the p-value is < 0.02 or > 0.50.

The acc^*_{eq} statistic creates a problem for significance testing because it optimizes a latent tie threshold for each metric on each test set (just one threshold for all item-wise score vectors). Since the permutation test for comparing two metrics creates two new vectors by randomly swapping elements of the original vectors on each draw, this necessitates the very expensive step of finding two new tie thresholds for each draw. To reduce the expense, we used the following approximate procedure. First find an optimal threshold for each input metric on the current test set, then create all pairs of item-wise scores and assign a correct/incorrect status to each pair by examining whether the metric's ranking matches the human ranking. Then perform the permutation test on these pairwise status vectors rather than the original score vectors. This approximation has more degrees of freedom than the original test, and can sample pairs that would never result from swapping the original score vectors, but our experiments showed that it is a reasonable proxy for the correct procedure.

To compute overall p-values based on weighted average scores of two metrics across all tasks, we cache the results of the draws for the per-task significance tests. In all cases, these are vectors of K pairs of correlation or accuracy statistics. Where $K < 1000$ due to early stopping, we duplicate elements to get 1000 examples. Then for i in 1..1000 we compare the weighted average of the pairs from the ith draw across all tasks, and record the results to produce an overall p-value.

Clustering

Given significance results (p-values) for all pairs of metrics, we assign ranks as follows. Starting with the highest-scoring metric, we move down the list of metrics in descending order by score, and assign rank 1 to all metrics until we encounter the first metric that is significantly different from any that have been visited so far. That metric is assigned rank 2, and the process is repeated. This continues until all metrics have been assigned a rank. Note that this is a greedy algorithm, and hence it can place two metrics that are statistically indistinguishable in different clusters.

6 Main Results

As we have seen in Section 5, the main results are the overall scores by taking a weighted average of the results from the ten main tasks, including system-level and segment-level tasks in different translation directions. Similar to last year, since the main use case of automatic metrics is to rank systems, system-level accuracy has a 1/4 weight on the final score with the remaining 3/4 distributed over 9 different settings.

Table 1 shows the official scores and rankings of all baselines and primary submissions. Table 8 and 9 show the scores and rankings of each individual task at system level and segment level, respectively. Similar to last year's results, neural metrics perform significantly better than lexical metrics. Of the 32 evaluated metrics, BLEU, F200SPBLEU and CHRF are ranked 28th, 24th and 29th respectively. On the other hand, fine-tuned neural baseline metrics, like COMET and BLEURT-20, remain ranked higher than several of the new primary submissions. They are surpassed only by submissions relying on significantly larger models.

It is worth noting that the best-performing baseline, COMETKIWI, along with four of the seven top-performing primary submissions, are reference-free. As we will elaborate on in a later section (Section 8), there are quality issues with human reference translations. This highlights the challenge of ensuring robustness to poor-quality references for reference-based metrics. In cases where a high-quality human reference is not available, reference-free metrics can serve as more robust alternatives.

Overall, XCOMET-Ensemble is the best performing metric in terms of average scores over the 10 meta-evaluation settings, with a statistically significant advantage over all other metrics. It consistently correlates best with human MQM scores at segment level for all translation directions, and it is ranked at worst in the 2nd significance cluster for all system-level meta-evaluation tasks.

Figure 1 shows the correlation scores split by translation direction. There are two key observations: 1) a majority of the metrics have higher correlations for en→de among the three translation directions, except for MRE-SCORE-LABSE-REGULAR and EBLEU, that perform substantially better for he→en, and YISI-1 and BERTSCORE, that perform equally in en→de and he→en; 2) reference-based metrics struggle for zh→en due to the reference quality, except for XCOMET-

Metric	avg-corr		en→de,he→en,zh→en accuracy task1		en→de pearson task2		he→en pearson task5		zh→en pearson task8	
XCOMET-Ensemble	**1**	**0.825**	**1**	**0.928**	2	0.980	**1**	**0.950**	2	0.927
XCOMET-QE-Ensemble*	2	0.808	**1**	**0.908**	2	0.974	2	0.909	3	0.892
MetricX-23	2	0.808	**1**	**0.908**	2	0.977	2	0.910	4	0.873
GEMBA-MQM*	2	0.802	**1**	**0.944**	**1**	**0.993**	2	0.939	**1**	**0.991**
MetricX-23-QE*	2	0.800	2	0.892	2	0.969	3	0.858	4	0.859
mbr-metricx-qe*	3	0.788	2	0.880	2	0.976	2	0.915	2	0.936
MaTESe	3	0.782	2	0.904	4	0.918	2	0.906	3	0.889
CometKiwi*	3	0.782	**1**	**0.904**	3	0.946	3	0.860	2	0.963
COMET	3	0.779	2	0.900	**1**	**0.990**	2	0.940	3	0.898
BLEURT-20	3	0.776	2	0.892	**1**	**0.990**	2	0.937	4	0.880
KG-BERTScore*	3	0.774	2	0.884	4	0.926	2	0.908	2	0.962
sescoreX	3	0.772	2	0.892	3	0.952	3	0.901	5	0.797
cometoid22-wmt22*	4	0.772	2	0.880	2	0.973	4	0.839	2	0.940
docWMT22CometDA	4	0.768	2	0.904	**1**	**0.990**	2	0.922	3	0.907
docWMT22CometKiwiDA*	4	0.767	2	0.900	2	0.970	2	0.906	2	0.965
Calibri-COMET22	4	0.767	**1**	**0.904**	2	0.963	2	0.930	4	0.863
Calibri-COMET22-QE*	4	0.755	2	0.863	2	0.978	4	0.778	2	0.934
YiSi-1	4	0.754	2	0.871	4	0.925	2	0.917	4	0.823
MS-COMET-QE-22*	5	0.744	2	0.871	3	0.959	5	0.721	3	0.901
prismRef	5	0.744	2	0.851	4	0.920	**1**	**0.956**	6	0.762
mre-score-labse-regular	5	0.743	2	0.888	3	0.942	**1**	**0.958**	3	0.903
BERTscore	5	0.742	2	0.871	5	0.891	3	0.895	5	0.810
XLsim	6	0.719	2	0.855	4	0.925	3	0.887	5	0.796
f200spBLEU	7	0.704	3	0.819	4	0.919	4	0.805	6	0.772
MEE4	7	0.704	3	0.823	5	0.861	3	0.879	6	0.743
tokengram_F	7	0.703	3	0.815	5	0.858	3	0.878	5	0.795
embed_llama	7	0.701	3	0.831	5	0.861	4	0.841	5	0.785
BLEU	7	0.696	3	0.815	4	0.917	5	0.769	7	0.734
chrF	7	0.694	3	0.795	5	0.866	4	0.776	5	0.809
eBLEU	7	0.692	2	0.859	4	0.918	2	0.911	7	0.727
Random-sysname*	8	0.529	4	0.578	6	0.357	6	0.209	8	0.093
prismSrc*	9	0.455	5	0.386	6	-0.327	6	-0.017	8	-0.406

Table 8: Results on system-level tasks for main language pairs. Rows are sorted by the overall average correlation across all 10 tasks (leftmost column). Starred metrics are reference-free, and underlined metrics are baselines.

ENSEMBLE and SESCOREX. The reason for the significant drop in correlation for he→en is unclear. This drop is observed across almost all metrics, whether they are trained or untrained, reference-free or reference-based, and they exhibit varying degrees of degradation.

We continue to be interested in metrics' ability to generalise across domains. In Figure 2, 3 and 4 we present the performance of each metric across different domains in each translation direction. Most metrics perform well in evaluating translation in the user reviews domain across translation direction, despite lacking annotated data in that domain. Further investigation is required to understand whether this is because the translation quality of MT output is more diverse in the user reviews domain, making it easier for metrics to accurately discriminate.

Figure 5 shows the average correlations of metrics when grouped separately by system-level and segment-level tasks. Many metrics fall into the same significance cluster when evaluated on the system-level, as we only have a limited number of MT systems. Although most of the metrics compute the system-level score by averaging their segment-level scores, we observe that high correlations between automatic metrics and human judgments at the segment level do not necessarily guarantee high correlations at the system level. For example, PRISMSRC is in the middle of the pack and has moderate Pearson's correlation at segment level for en→de. However, it is negatively correlating with human judgements when evaluating the same language pair at system level.

7 Understanding metrics' scores beyond correlation

In the past few years, we demonstrated that new metrics correlate better with human judgments than BLEU does. Some new baseline metrics even consistently outperform BLEU for consecutive years across translation directions and domains. How-

Metric	en→de pearson task3		en→de acc-t task4		he→en pearson task6		he→en acc-t task7		zh→en pearson task9		zh→en acc-t task10	
XCOMET-Ensemble	**1**	**0.695**	**1**	**0.604**	**1**	**0.556**	**1**	**0.586**	**1**	**0.650**	**1**	**0.543**
XCOMET-QE-Ensemble*	2	0.679	3	0.588	3	0.498	4	0.554	**1**	**0.647**	3	0.533
MetricX-23	4	0.585	**1**	**0.603**	**1**	**0.548**	2	0.577	2	0.625	3	0.531
GEMBA-MQM*	6	0.502	5	0.572	5	0.401	3	0.564	6	0.449	5	0.522
MetricX-23-QE*	3	0.626	2	0.596	2	0.520	3	0.564	**1**	**0.647**	4	0.527
mbr-metricx-qe*	4	0.571	3	0.584	5	0.411	4	0.553	5	0.489	2	0.537
MaTESe	5	0.554	9	0.528	4	0.459	5	0.550	4	0.511	12	0.479
CometKiwi*	7	0.475	5	0.569	7	0.387	6	0.544	6	0.442	4	0.525
COMET	8	0.432	4	0.574	5	0.401	8	0.532	8	0.396	7	0.514
BLEURT-20	7	0.484	5	0.572	8	0.382	10	0.519	9	0.378	6	0.518
KG-BERTScore*	8	0.451	7	0.556	8	0.382	7	0.537	7	0.430	6	0.516
sescoreX	5	0.519	6	0.563	7	0.385	15	0.484	3	0.536	9	0.499
cometoid22-wmt22*	8	0.441	4	0.578	9	0.365	11	0.515	5	0.479	7	0.515
docWMT22CometDA	10	0.394	7	0.559	10	0.339	13	0.497	10	0.353	10	0.493
docWMT22CometKiwiDA*	8	0.444	8	0.547	12	0.286	14	0.489	8	0.387	10	0.493
Calibri-COMET22	9	0.413	10	0.522	5	0.401	11	0.515	8	0.396	14	0.474
Calibri-COMET22-QE*	8	0.441	12	0.483	6	0.395	12	0.506	6	0.443	10	0.491
YiSi-1	11	0.366	8	0.542	6	0.395	8	0.529	11	0.290	8	0.504
MS-COMET-QE-22*	12	0.310	8	0.546	12	0.295	13	0.498	9	0.367	9	0.498
prismRef	6	0.516	10	0.518	11	0.319	9	0.528	14	0.183	8	0.504
mre-score-labse-regular	17	0.111	9	0.530	8	0.378	10	0.522	16	0.145	12	0.481
BERTscore	12	0.325	9	0.528	10	0.335	11	0.515	12	0.236	9	0.499
XLsim	13	0.239	9	0.527	14	0.233	16	0.480	17	0.111	15	0.464
f200spBLEU	14	0.237	9	0.526	14	0.230	18	0.447	18	0.108	13	0.476
MEE4	16	0.202	9	0.529	13	0.256	19	0.441	18	0.105	12	0.480
tokengram_F	15	0.227	10	0.520	14	0.226	17	0.461	20	0.060	11	0.485
embed_llama	13	0.250	12	0.483	15	0.215	20	0.430	15	0.161	16	0.447
BLEU	16	0.192	10	0.520	15	0.220	19	0.442	17	0.119	14	0.472
chrF	14	0.232	10	0.519	15	0.221	17	0.460	19	0.063	11	0.485
eBLEU	19	-0.011	11	0.512	16	0.131	18	0.445	22	-0.084	14	0.473
Random-sysname*	18	0.064	14	0.409	17	0.041	20	0.428	21	0.018	18	0.381
prismSrc*	9	0.425	13	0.426	16	0.140	19	0.441	13	0.223	17	0.421

Table 9: Results on segment-level tasks for main language pairs. Rows are sorted by the overall average correlation across all 10 tasks (leftmost column in Table 8). Starred metrics are reference-free, and underlined metrics are baselines.

ever, the research community is still reluctant to adopt newer and better automatic MT evaluation metrics in practice. One of the reasons is that MT researchers have established some "common beliefs" about the relationship between BLEU and actual translation quality, and similar intuitions about new metrics have yet to crystallize. Thus, this year, we conduct two additional analyses beyond correlation with human to understand the meaning of the score differences that metrics present with respect to the statistical significance of MT system rankings according to human annotations and metric scores. Our results should *NOT* be used as arguments to forego significance tests or appropriate human evaluation. These analyses only support an intuitive sense of metric score meanings to encourage broader adoption of new automatic MT evaluation metrics.

7.1 Correspondence to MQM scores significance

First, we follow Lo et al. (2023a) to study the relationship between statistically significant differences in human scores and the magnitude of metric differences. Specifically, we run a one-sided paired t-test with an equal variance assumption for each system pair on segment-level MQM scores. After that, we fit the corresponding metric score differences and the p-values of the t-test on the MQM scores to an isotonic regression (Robertson et al., 1988), that predicts whether the human MQM score difference will be significant given the metric's score difference. Isotonic regression produces a non-decreasing function where the classifier output can be interpreted as a confidence level.[9] We set $p_{mqm} < 0.05$ as the significance level of MQM

[9] https://scikit-learn.org/stable/modules/isotonic.html

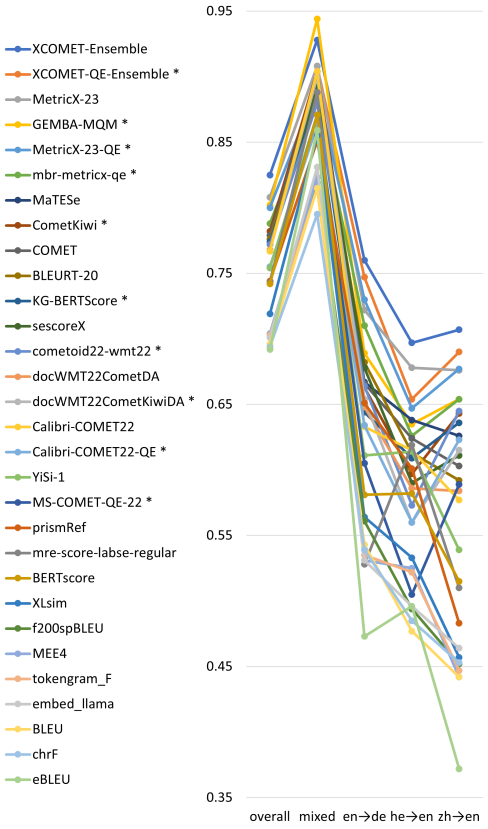


Figure 1: Average metrics' meta-evaluation scores in tasks grouped by translation direction. The "mixed" group is the accuracy score of the metrics in task 1.

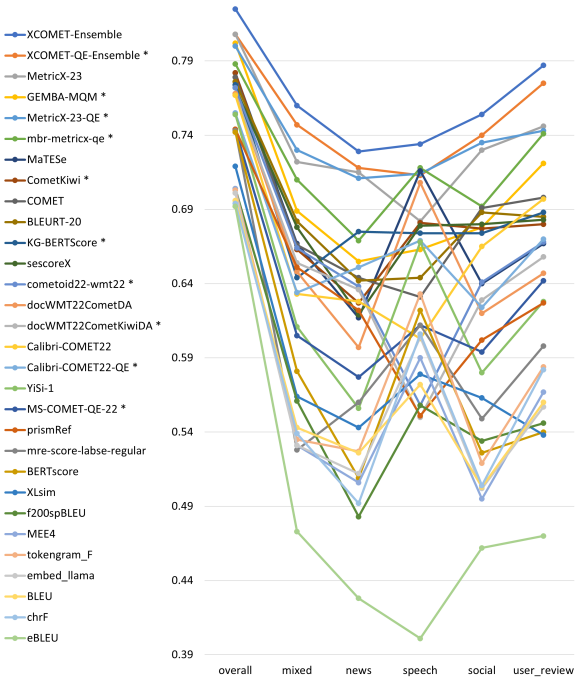


Figure 2: Average metrics' correlation with human in tasks grouped by domain in en→de. The "mixed" group is the average correlation in all en→de tasks.

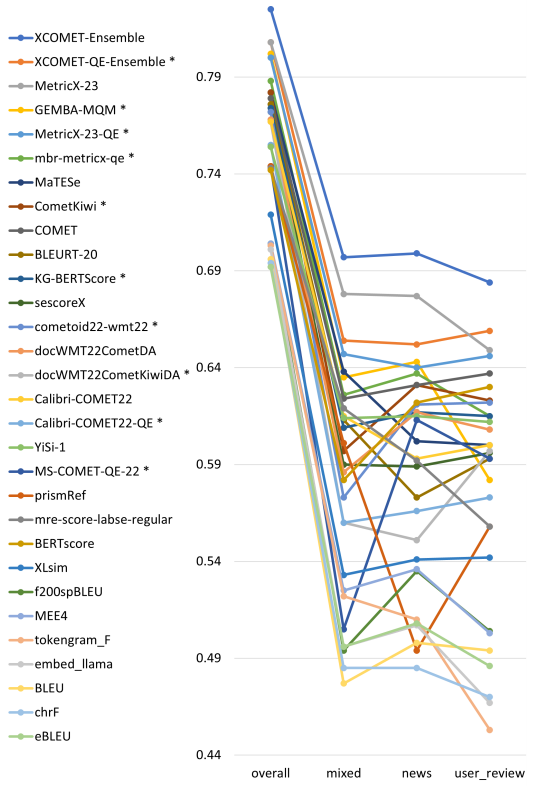


Figure 3: Average metrics' correlation with human in tasks grouped by domain in he→en. The "mixed" group is the average correlation in all he→en tasks.

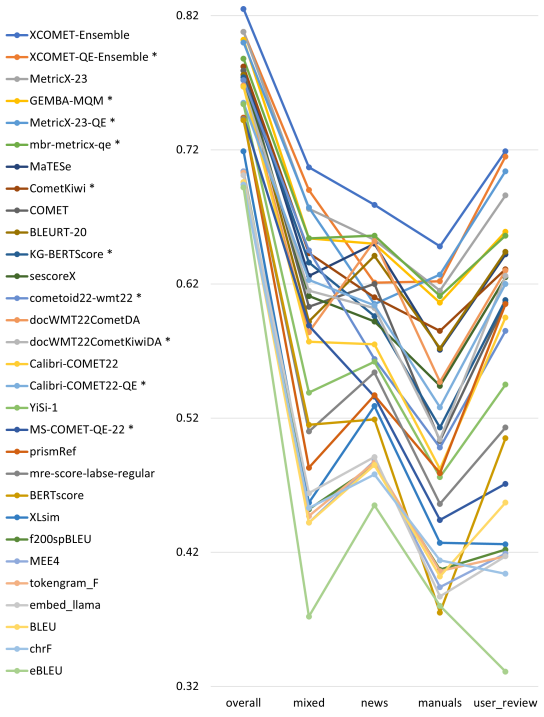


Figure 4: Average metrics' correlation with human in tasks grouped by domain in zh→en. The "mixed" group is the average correlation in all zh→en tasks.

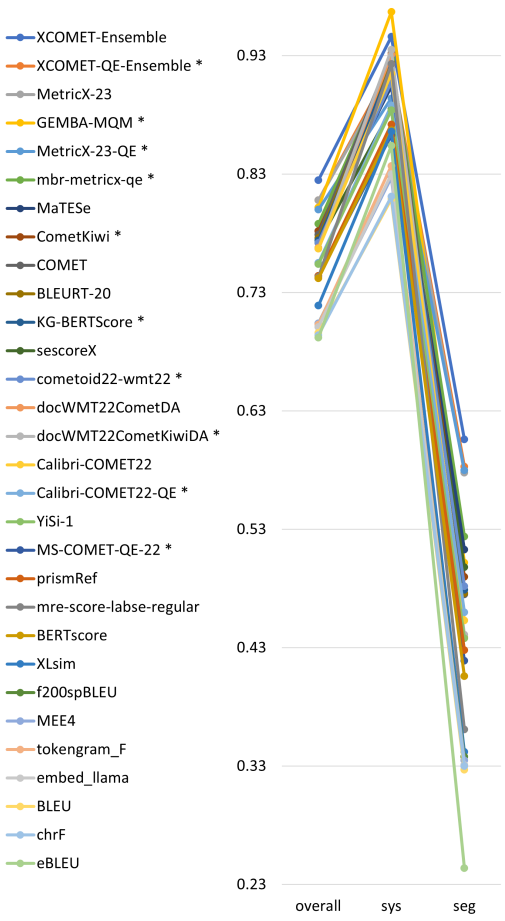


Figure 5: Average metrics' correlation with human in tasks grouped by granularity level.

scores. Thus, the output of the isotonic regression function can be viewed as $Pr(p_{mqm} < 0.05|\Delta M)$ where p_{mqm} is the p-value of the t-test on the MQM scores for each system pair and ΔM is the metric score difference.

Figure 6 shows the (log) p-value of one-sided paired t-test on the MQM scores against the corresponding BLEU and COMET score difference for each system pair in en→de. Figures 9-14 in appendix D, show the same analyses for all metrics and translation directions. For each metric, we can choose a particular level of confidence (i.e., a point along the y-axis on the right) to give metric score difference cut-offs (i.e., a point along the x-axis) that this metric difference reflects significant MQM score differences. Drawing a horizontal line from the confidence level, say 80%, to the red line enables us to find the minimum metric difference cut-off required at the corresponding x-value down from the red line, i.e. 11 for BLEU in Figure 6. Using this lookup method, Table 10 shows the cut-offs of ΔM when $Pr(p_{mqm} < 0.05|\Delta M) = 0.8$ for each metric and translation directions.

We run the leave-one-system-out cross validation and Table 10 shows that the range of precision in the cross validation are consistently high across metrics, with the exception of BLEU, CHRF, PRISMSRC, RANDOM-SYSNAME and SLIDE. This means the metric cut-offs we find using the regression model are reliable.

Contrary to the common belief that 2 BLEU improvement represents "significant" or "notable by human" improvement in the actual translation quality, our analyses show that 2.2 BLEU is the minimum required improvement for a high confidence (80%) that MQM annotators to mark significant differences in the translation output for one translation direction (zh→en) and that threshold would be as high as 11 BLEU for en→de. Table 10 serves as a reference between BLEU differences and differences in some of the modern metrics, and assists metric users in understanding scores provided by modern metrics. For example, when evaluating he→en translation quality, we see that a BLEU difference of 3.5 corresponds to 80% confidence that the metric's ranking of the two MT systems will match with the decision made by human annotators with a significant difference. Meanwhile, a COMET score difference of 0.014 would have the same 80% chance of human judged significant difference.

7.2 Correspondence to metric scores significance

Inspired by Marie (2022), we run a study similar to that in the previous subsection but on the relations between statistically significant differences in metric scores and the magnitude of metric differences. Instead of one-sided t-test on MQM, the p-values are now obtained by running statistical significance tests with bootstrap resampling on the metric scores for each system pair. Similarly, we fit the corresponding metric score differences and the p-values of the significance test to an isotonic regression for predicting whether the translation quality improvement as indicated by the metric will be significant given the metric score difference. We set $p_M < 0.05$ and thus, the output of the isotonic regression function is now $Pr(p_M < 0.05|\Delta M)$, where p_M is the p-value of the significance test on the metric scores for each system pair and ΔM is the metric score difference.

Figure 7 shows the (log) p-value of the signifi-

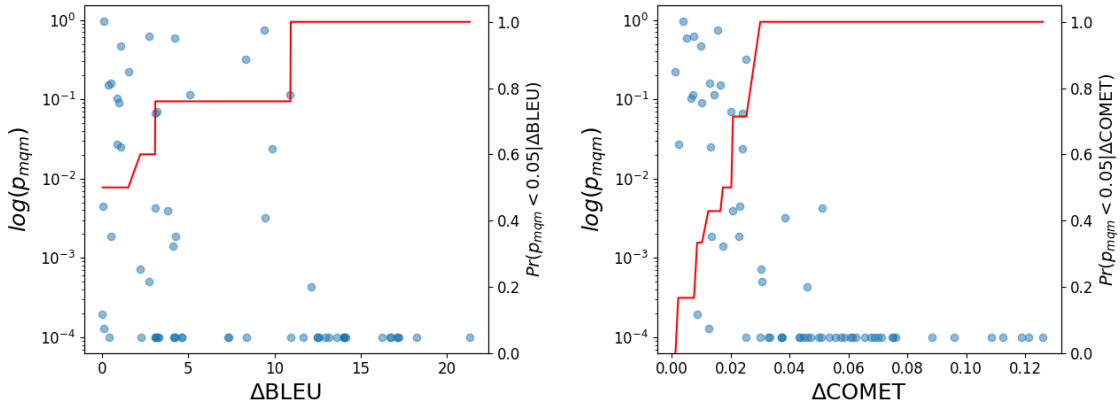


Figure 6: Log p-value of one-sided paired t-test on MQM scores (p_{mqm}) against the metric (left: BLEU, right: COMET) score difference for each system pair in en→de. The red line is the isotonic regression fit to all data points, representing $Pr(p_{mqm} < 0.05|\Delta M)$. Note: for readability, values of p_{mqm} are rounded up to 0.0001 when they are less than 0.0001.

	en→de		he→en		zh→en	
Metric	min ΔM	c.v. precision	min ΔM	c.v. precision	min ΔM	c.v. precision
BERTSCORE	0.011	[75-100%]	0.0053	[83-100%]	0.0033	[75-100%]
BLEU	11	[33-100%]	3.5	[82-100%]	2.2	[75-100%]
BLEURT-20	0.041	[75-100%]	0.019	[100-100%]	0.013	[82-100%]
CALIBRI-COMET22	0.068	[71-100%]	0.031	[89-100%]	0.043	[80-100%]
CALIBRI-COMET22-QE	0.072	[82-100%]	0.020	[86-100%]	0.025	[67-100%]
CHRF	2.8	[25-100%]	3.2	[83-100%]	2.6	[86-100%]
COMET	0.030	[78-100%]	0.014	[88-100%]	0.013	[80-100%]
COMETKIWI	0.022	[67-100%]	0.014	[64-100%]	0.0098	[62-100%]
COMETOID22-WMT22	0.018	[86-100%]	0.0077	[71-100%]	0.011	[67-100%]
DOCWMT22COMETDA	0.027	[78-100%]	0.012	[82-100%]	0.014	[82-100%]
DOCWMT22COMETKIWIDA	0.026	[75-100%]	0.012	[64-100%]	0.0096	[71-100%]
EBLEU	0.022	[57-100%]	0.019	[83-100%]	0.017	[86-100%]
EMBED_LLAMA	0.062	[67-100%]	0.019	[80-100%]	0.020	[80-100%]
F200SPBLEU	4.6	[60-100%]	3.6	[75-100%]	3.5	[86-100%]
GEMBA-MQM	2.0	[89-100%]	1.0	[82-100%]	2.0	[69-100%]
KG-BERTSCORE	0.0097	[50-100%]	0.0097	[86-100%]	0.0079	[62-100%]
MATESE	0.99	[71-100%]	0.77	[75-100%]	0.70	[73-100%]
MBR-METRICX-QE	0.047	[75-100%]	0.026	[82-100%]	0.022	[75-100%]
MEE4	0.013	[71-100%]	0.024	[78-100%]	0.020	[86-100%]
METRICX-23	0.73	[100-100%]	0.29	[76-100%]	0.55	[83-100%]
METRICX-23-QE	0.53	[71-100%]	0.092	[67-100%]	0.49	[60-100%]
MRE-SCORE-LABSE-REGULAR	0.010	[67-100%]	0.016	[100-100%]	0.0064	[62-100%]
MS-COMET-QE-22	1.5	[80-100%]	1.4	[67-100%]	1.2	[60-100%]
PRISMREF	0.081	[75-100%]	0.14	[88-100%]	0.19	[83-100%]
PRISMSRC	0.036	[73-100%]	0.040	[33-100%]	0.022	[64-100%]
RANDOM-SYSNAME	7.8	[0-100%]	0.082	[67-90%]	5.0	[50-90%]
SESCOREX	0.38	[73-100%]	0.50	[89-100%]	0.62	[73-100%]
SLIDE	0.049	[78-100%]	0.017	[78-100%]	0.013	[58-100%]
XCOMET-ENSEMBLE	0.029	[88-100%]	0.0092	[83-100%]	0.012	[75-100%]
XCOMET-QE-ENSEMBLE	0.038	[86-100%]	0.012	[83-100%]	0.021	[67-100%]
XLSIM	0.015	[67-100%]	0.0073	[82100%]	0.0091	[70-100%]
YISI-1	0.0049	[67-100%]	0.0060	[80-100%]	0.0054	[75-100%]

Table 10: Minimum ΔM when $Pr(p_{mqm} < 0.05|\Delta M) = 0.8$ for each metric in different translation directions round to 2 significant figures, and the range of precision for the isotonic regression model in leave-one-system-out cross validation.

cance test with bootstrap resampling on the metric scores for BLEU and COMET score difference of each system pair in en→de. Additional figures (Figures 15-20 in appendix Appendix D) show the same analyses for all metrics and translation directions. Using the same lookup method described in the previous subsection, Table 11 shows the cut-offs of ΔM when $Pr(p_M < 0.05|\Delta M) = 0.8$ for each metric and translation directions.

We run the leave-one-system-out cross valida-

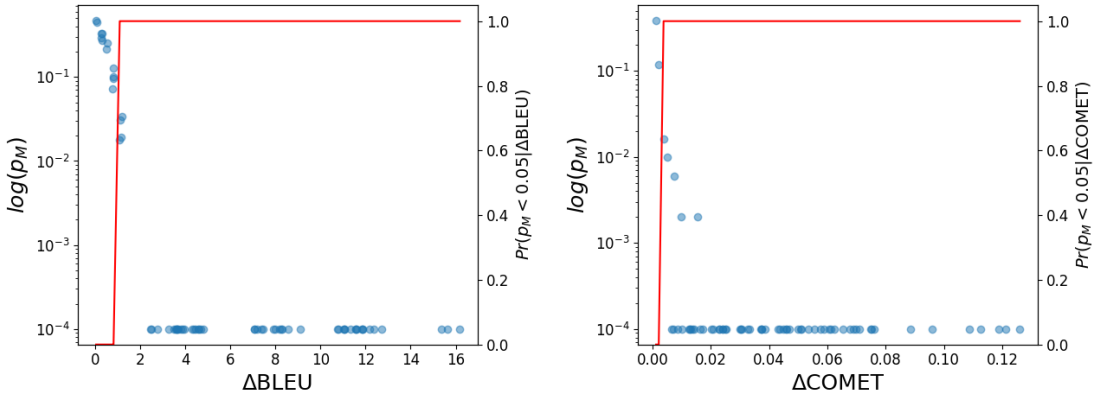


Figure 7: Log p-value of significance test with bootstrap resampling (p_M) on system-level metric scores against each metric (left: BLEU, right: COMET) score difference for each system pair in en→de. The red line is the isotonic regression fit to all data points, representing $Pr(p_M < 0.05|\Delta M)$. Note: for readability, values of p_M are rounded up to 0.0001 when they are less than 0.0001.

	en→de		he→en		zh→en	
Metric	min ΔM	c.v. precision	min ΔM	c.v. precision	min ΔM	c.v. precision
BERTSCORE	0.0026	[100-100%]	0.0012	[100-100%]	0.00085	[100-100%]
BLEU	1.1	[100-100%]	0.79	[100-100%]	0.58	[93-100%]
BLEURT-20	0.0081	[100-100%]	0.0041	[100-100%]	0.0024	[100-100%]
CALIBRI-COMET22	0.010	[91-100%]	0.0063	[100-100%]	0.0064	[100-100%]
CALIBRI-COMET22-QE	0.015	[100-100%]	0.0086	[89-100%]	0.0078	[92-100%]
CHRF	0.99	[100-100%]	0.68	[100-100%]	0.48	[100-100%]
COMET	0.0038	[100-100%]	0.0038	[90-100%]	0.0029	[100-100%]
COMETKIWI	0.0074	[91-100%]	0.0019	[100-100%]	0.0025	[93-100%]
COMETOID22-WMT22	0.0062	[82-100%]	0.0026	[100-100%]	0.0019	[100-100%]
DOCWMT22COMETDA	0.0033	[100-100%]	0.0013	[100-100%]	0.0023	[100-100%]
DOCWMT22COMETKIWIDA	0.0028	[100-100%]	0.0021	[100-100%]	0.0015	[100-100%]
EBLEU	0.0076	[90-100%]	0.0048	[100-100%]	0.0050	[100-100%]
EMBED_LLAMA	0.013	[100-100%]	0.0079	[100-100%]	0.0054	[100-100%]
F200SPBLEU	1.0	[100-100%]	0.94	[100-100%]	0.65	[100-100%]
GEMBA-MQM	0.52	[100-100%]	0.38	[100-100%]	0.35	[100-100%]
KG-BERTSCORE	0.0051	[100-100%]	0.0016	[100-100%]	0.00029	[93-100%]
MATESE	0.33	[100-100%]	0.20	[100-100%]	0.15	[100-100%]
MBR-METRICX-QE	0.0073	[100-100%]	0.0039	[100-100%]	0.0023	[100-100%]
MEE4	0.0029	[90-100%]	0.0067	[100-100%]	0.0054	[100-100%]
METRICX-23	0.23	[100-100%]	0.083	[90-100%]	0.089	[92-100%]
METRICX-23-QE	0.19	[100-100%]	0.072	[89-100%]	0.11	[100-100%]
MRE-SCORE-LABSE-REGULAR	0.0034	[100-100%]	0.0028	[100-100%]	0.0010	[100-100%]
MS-COMET-QE-22	0.49	[100-100%]	0.45	[88-100%]	0.18	[100-100%]
PRISMREF	0.018	[100-100%]	0.031	[100-100%]	0.020	[100-100%]
PRISMSRC	0.028	[100-100%]	0.025	[75-100%]	0.016	[100-100%]
RANDOM-SYSNAME	0.21	[100-100%]	0.14	[100-100%]	0.12	[100-100%]
SESCOREX	0.039	[100-100%]	0.10	[100-100%]	0.085	[100-100%]
XCOMET-ENSEMBLE	0.010	[90-100%]	0.0035	[100-100%]	0.0033	[100-100%]
XCOMET-QE-ENSEMBLE	0.0065	[100-100%]	0.0027	[100-100%]	0.0042	[93-100%]
XLSIM	0.0019	[100-100%]	0.0018	[100-100%]	0.0022	[100-100%]
YISI-1	0.0013	[100-100%]	0.0033	[73-100%]	0.00074	[100-100%]

Table 11: Minimum ΔM when $Pr(p_M < 0.05|\Delta M) = 0.8$ for each metric in different translation directions round to 2 significant figures, and the range of precision for the isotonic regression model in leave-one-system-out cross validation.

tion, and Table 11 shows that the range of precision in the cross validation are consistently high across metrics. This means the metric cut-offs we find using the regression model are reliable.

Our results, agreeing with Marie (2022), show that to claim significant differences ($p_M < 0.05$) in BLEU with high confidence (80%), the BLEU differences should be greater than 1.1 BLEU for en→de. Table 11 serves as a reference of metric differences with respect to statistical significance with high confidence. For example, when evaluating en→de translation quality, we see that a BLEU

difference of 1.1 corresponds to 80% confidence the difference is statistical significant. Meanwhile, a COMET score difference of 0.0038 would have the same 80% chance of statistical significance.

We have to emphasize again that our result should *NOT* be interpreted as evidence to forego significance tests or appropriate human evaluation. Instead, we are only providing assistance to build an intuition on the meaning of the scores provided by the new metrics to encourage the transition away from BLEU.

8 Synthetic Reference Translations

Reference-based metrics compare machine translations of source segments to human translations of those same source segments to determine how good they are. The quality of the underlying human translation is crucial and can impact the quality of the predicted scores more than the choice of metric (Freitag et al., 2020). Motivated by the low human ratings of refA for Chinese→English (Table 4) and the relatively high rankings of reference-free metrics (in comparison to other language pairs) for this language-pair, we investigate a method for generating a synthetic reference translation based on the MT output and the corresponding MQM ratings.

8.1 Synthetic Reference Generation

The main idea is straightforward: Given the set of translations of WMT23 General MT Shared Task (generalMT2023) from the WMT campaign and their corresponding MQM ratings, we generate a new synthetic reference translation by choosing for each segment the translation that received the lowest MQM error score as the selected reference. The original human reference translation (i.e. refA) is considered as one of the possible translations in this process, and MQM score ties are broken randomly. Table 12 shows the resulting MQM score of the synthetic reference translations. We were able to reduce the MQM score to below 1 for both tested language pairs (en→de and zh→en), which corresponds to an average of less than one minor error per segment. While this may seem like a significant improvement, we must caution the reader that this is in essence "cherry-picking" based on the MQM ratings and may therefore introduce many hidden issues.

It is also interesting to understand how many segments come from each of the individual MT systems in this selection process. Table 13 shows the number of segments contributed by each system to the generated synthetic reference translations. Unsurprisingly, the top performing MT systems are also the main contributors to the selected synthetic reference translation. For en→de, refA (the original human-generated reference translation) provided the majority of the selected translations, while for zh→en GPT4-5shot is the main contributor, reflecting that the human-generated reference refA for zh→en was indeed error-prone. However, it is interesting to note that despite the overall low quality of this human-generated reference, our method still selected 209 segments from this translation as the lowest-error translation. This would appear to indicate that these human-generated reference translations are not uniformly bad, and only a subset of the translations were unreliable and contained major errors. A possible explanation could be that multiple translators worked on the reference, however, we confirmed with the sponsor translating zh→en that all segments were translated with the same translator.

	zh→en	en→de
synthetic Ref.	0.66	0.87
best MT	2.10	3.72
refA	4.83	2.96

Table 12: MQM scores of the synthetic references.

zh→en		en→de	
GPT4-5shot	314	refA	243
refA	209	GPT4-5shot	57
Lan-BridgeMT	157	ONLINE-B	36
ANVITA	142	ONLINE-A	20
HW-TSC	105	AIRC	20
IOL_Research	42	ONLINE-W	19
ONLINE-W	33	NLLB_Greedy	14
ONLINE-Y	28	NLLB_MBR_BLEU	13
ONLINE-B	26	ONLINE-G	10
ONLINE-A	24	ONLINE-Y	9
ONLINE-G	21	Lan-BridgeMT	9
NLLB_Greedy	20	ONLINE-M	8
ZengHuiMT	18	ZengHuiMT	2
Yishu	18		
NLLB_MBR_BLEU	14		
ONLINE-M	6		

Table 13: Number of segments contributed by each system towards the synthetic reference.

8.2 Impact on Metrics

Figure 8 compares the segment-level and system-level Pearson correlations of all submitted metrics

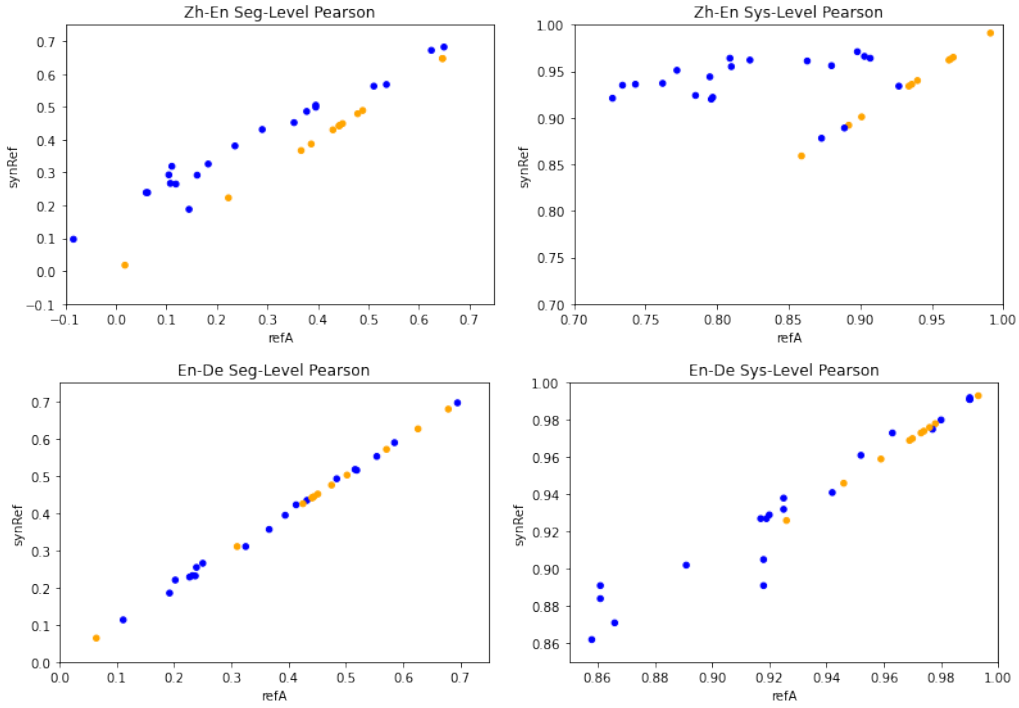


Figure 8: Pearson Correlation when using either the synthetic ref or the original human translation as reference translation. QE metrics are coloured in yellow; ref-based metrics are coloured in blue.

when using either the original or the synthetic reference translation. Reference-based metrics are coloured in blue, while QE metrics are coloured in orange. Obviously, as QE metrics do not use reference translations, their correlations are exactly the same. For Chinese→English, replacing the human-generated reference translation by the synthetic reference translation has a dramatic impact. All reference-based metrics increased their correlation levels with human judgements at both the segment-level and the system-level. This clearly indicates how critically important a high quality reference translation is for reference-based metrics, but moreover, it also highlights the advantages of QE metrics in cases where human-generated references have major quality issues. For the English→German language pair, the human-generated reference translation is of higher quality than any submitted MT system. Consequently, the synthetic reference translation had almost no impact on the segment-level correlations and only a mixed impact on the system-level correlations.

The main takeaways from this study are (i) poor human-generated reference translations can dramatically hurt the performance and reliability of your metric, (ii) strong QE metrics can be better alternatives in such scenarios, and (iii) generating a synthetic reference translation from all system outputs can be used to mitigate bad reference translations, although it assumes obtaining MQM annotations and suffers from cherry-picking bias.[10]

An open unanswered question remains: is it always necessary for a reference translation to be of higher quality than the translation generated by the MT system, in order to have a reliable reference-based metric? This would imply that generating a synthetic reference translation with any errors is problematic, since for any reference-based automatic metric, these synthetic references would become useless for evaluating any MT system that generates translations that surpasses the reference in quality.

9 DA+SQM Human Evaluation

In addition to our MQM annotations and as a contrastive evaluation to cover more language pairs, we look into the performance of metrics when compared to the human evaluation campaign conducted by the WMT23 General MT Shared Task (Kocmi et al., 2023), who ran human evaluation for all 14 translation directions and all WMT23 submissions.

In contrast to previous years, they no longer use

[10]Among other issues, any practical strategy for creating synthetic references would need to have a way of avoiding bias toward systems that are similar to the ones used for reference creation.

MTurk neither reference-based evaluation for into-English language pairs. They no longer use z-score normalization because the user interface decision to not track users (i.e., only maintaining HIT information) means that the z-scores are likely to be influenced by the distribution of system quality in the HITs rather than only annotator variation.

They employ the Direct Assessment Scalar Quality Metrics (DA+SQM) technique as presented in Kocmi et al. (2022a).

DA+SQM asks bilingual raters to annotate system translations against original sources on a 0–100 labelled scale. The scale is marked with seven points representing expected quality.

At the time of writing, the WMT23 General MT Shared Task had collected data only for 8 translation directions: Chinese↔English (zh↔en), German↔English (de↔en), Japanese↔English (ja↔en), English→Czech (en→cz), and Czech→Ukrainian (cz→uk).

We present system-level accuracy results for both MQM and DA+SQM in Table 14. There are many factors that could affect the ranking. Apart from using a different human annotation protocol, MQM compares 3 translation directions whereas the DA+SQM compares 8 translation directions, containing also the non-English low-resource pair of cz→uk. There is an overlap of only two translation directions between the two: en→de and zh→en. The main difference in ranking is for metrics XCOMET-Ensemble and MetricX-23 ranking significantly lower than for MQM. Investigating system-level Pearson's correlation for individual languages in Tables 19 to 27 shows that both metrics are performing considerably lower across all languages (except en→cz and cz→uk) and we do not see any pattern behind the drop in performance.

Metric	MQM		DA+SQM	
Translation directions	3		8	
System pairs (N)	237		793	
GEMBA-MQM*	0.944	(1)	0.899	(1)
XCOMET-Ensemble	0.928	(2)	0.870	(10)
MetricX-23	0.908	(3)	0.863	(11)
XCOMET-QE-Ensemble*	0.908	(4)	0.871	(8)
CometKiwi*	0.904	(5)	0.887	(3)
COMET	0.900	(6)	0.890	(2)
BLEURT-20	0.892	(7)	0.880	(6)
MetricX-23-QE*	0.892	(8)	0.870	(9)
mre-score-labse-regular	0.888	(9)	0.861	(12)
KG-BERTScore*	0.884	(10)	0.884	(4)
cometoid22-wmt22*	0.880	(11)	0.884	(5)
BERTscore	0.871	(12)	0.799	(16)
MS-COMET-QE-22*	0.871	(13)	0.879	(7)
YiSi-1	0.871	(14)	0.832	(13)
eBLEU	0.859	(15)	0.781	(19)
XLsim	0.855	(16)	0.831	(14)
prismRef	0.851	(17)	0.808	(15)
embed_llama	0.831	(18)	0.778	(20)
f200spBLEU	0.819	(19)	0.786	(17)
BLEU	0.815	(20)	0.770	(22)
tokengram_F	0.815	(21)	0.786	(18)
chrF	0.795	(22)	0.777	(21)
Random-sysname*	0.578	(23)	0.580	(23)
prismSrc*	0.386	(24)	0.412	(24)

Table 14: Comparison between system-level pairwise accuracy using MQM and DA+SQM gold scores. MQM results pool data from our 3 main language pairs; DA+SQM results pool data from the 8 language pairs for which DA+SQM scores are available. Rows are sorted by MQM accuracy, with the pure rank order indicated in brackets. Starred metrics are reference-free and underlined metrics are baselines.

10 Challenge Sets Sub-task

For the second year, we included a sub-task on challenge sets. This sub-task is inspired by the *Build it or break it: The Language Edition* shared task (Ettinger et al., 2017) which aimed at testing the generalizability of NLP systems beyond the distributions of their training data. Whereas the standard evaluation of the shared task runs on test sets containing generic text from real-world content, the challenge set evaluation is based on test sets designed with the aim of revealing the abilities or the weaknesses of the metrics on evaluating particular translation phenomena. In order to shed light on different perspectives on evaluation, the sub-task takes place in a decentralized manner, where, contrary to the main metric task, the test sets are not provided by the organizers but by different research teams, who are also responsible for analysing and presenting the results.

This subtask is made of three consecutive phases; 1) the *Breaking Round*, 2) the *Scoring Round* and 3) the *Analysis Round*:

1. In the *Breaking Round*, every challenge set participant (*Breaker*) submits their challenge set S composed of contrastive examples for different phenomena, where every example $(s, \hat{t}, t, r) \in S$ contains one source sentence s, one incorrect translation $\hat{t}$, one correct translation t and one reference r.

2. In the *Scoring Round*, the organizers decompose the S into a blind test set S', where each example includes either an incorrect translation $(s, \hat{t}, r)$ or a correct translation (s, t, r) along with the source and the reference. The separated contrastive examples are shuffled, and the golden truth of which samples are correct or incorrect is kept in a separate set. The

challenge set	directions	phenomena	items	citation	availability (https://github.com/)
ACES	146	translation errors	36476	Amrhein et al. (2023)	EdinburghNLP/ACES
DFKI-CS	3	linguistic phenomena	20993	Avramidis et al. (2023)	DFKI-NLP/mt-testsuite
MSLC23	4	low quality MT	9345	Lo et al. (2023b)	nrc-cnrc/MSLC23

Table 15: Overview of the participation at the metrics challenge sets sub-task

metrics participants from the main task (the *Builders*) are asked to score with their metrics the translations in the given blind test set without knowing which ones are correct or incorrect. Also, in this phase, the organizers score all data with the baseline metrics.

3. Finally, after having gathered all metric scores, the organizers return the respective scored translations to the *Breakers* for the *Analysis round*, where they look at which metrics are able to correctly rank the correct translations higher than the incorrect ones for the phenomena being tested.

There were 3 submissions this year, covering a wide range of phenomena and 146 different translation directions. An overview of the submitted challenge sets can be seen in Table 15. A short description of every submission follows:

ACES Challenge Set The Translation Accuracy ChallengE Set (ACES, Amrhein et al., 2023) consists of 36K examples representing challenges from 68 phenomena and covering 146 translation directions. The phenomena range from simple perturbations at the word/character level to more complex errors based on discourse and real-world knowledge. We benchmark the performance of segment-level metrics submitted to WMT 2023 using ACES. For each metric, the authors provide a detailed profile of performance across the ten top-level accuracy error categories in ACES as well as an overall *ACES-Score* for quick comparison. They also measure the incremental performance of the metrics submitted to both WMT 2023 and 2022.

They find that:

- there is no clear *winner* among the metrics submitted to WMT 2023,
- neural metrics also tend to focus more on lexical overlap than semantic content,
- reference-free metrics using language-agnostic multilingual embeddings struggle with detecting untranslated or sentences translated in the wrong direction, and
- performance change between the 2023 and 2022 versions of the metrics is highly variable.

The authors' recommendations are similar to those from WMT 2022. Metric developers should focus on: building ensembles of metrics from different design families, developing metrics that pay more attention to the source and rely less on surface-level overlap, and carefully determining the influence of multilingual embeddings on MT evaluation.

DFKI Challenge Set The submission by DFKI (Avramidis et al., 2023) employs a linguistically motivated challenge set that includes about 21,000 items extracted from 155 machine translation systems for three language directions (de→en, en→de, en→ru), covering more than 100 linguistically-motivated phenomena organized in 14 categories. The metrics that have the best performance with regard to our linguistically motivated analysis are the COMETOID22-WMT23 for de→en and METRICX-23-C for en→de and en→ru. Some of the most difficult phenomena for the metrics to score are *passive voice* for de→en, *named entities, terminology* and *measurement units* for en→de and *focus particles, adverbial clause* and *stripping* for en→ru.

MSLC23 Challenge Set The Metric Score Landscape Challenge (MSLC23; Lo et al., 2023b) data set aims to gain insight into metric scores on a broader/wider landscape of MT quality. Recent development of MT evaluation metrics has focused on improving their correlation with human judgment on translations of high-quality systems (e.g., participants in the WMT News/General MT Shared Tasks). This means that metric performance may be untested on low- to medium-quality MT output. MSLC23 provides a collection of low- to medium-quality MT output on the news portion of the WMT23 General MT Shared Task test set. Together with the high quality systems submitted to the General MT Shared Task, this enables better interpretation of metric scores across a range of different levels of translation quality. With this

wider range of MT quality, the authors also visualize and analyse metric characteristics beyond just correlation.

The authors find that the smaller variations in segment-level scores given by some metrics at the low end of quality could indicate that these metrics struggle to discriminate between low-quality MT systems. This is further shown by the observation that some metrics rank the low-quality systems in reverse order at system level. A "universal score" phenomenon for some metrics, where a small subset of non-minimum/maximum distinct scores are assigned to a variety of translation output, has been discovered. There is also an observation of diverse behaviours from different metrics on empty string translation. These results highlight the need for metric researchers to check their metrics' performance on a wider landscape of translation quality, or to indicate to potential users that they should be cautious about using their metric on a wide range of quality.

11 Conclusion

This paper summarizes the results of the WMT23 shared task on automated machine translation evaluation, the Metrics Shared Task. We presented an extensive analysis on how well metrics perform on our three main translation directions: English→German, Hebrew→English and Chinese→English. The results, based on 10 different tasks, confirm the superiority of neural-based learned metrics over overlap-based metrics like BLEU, SPBLEU or CHRF. These results are confirmed with DA+SQM human judgement. We also found that reference-free metrics were strong contenders this year, partly because they do not rely on the quality of reference translations, an increasingly important issue as MT systems under evaluation become better. In addition, we continued the challenge set subtask, where participants had to create contrastive test suites for evaluating metrics' ability to capture and penalise specific types of translation errors.

12 Ethical Considerations

MQM annotations and additional reference translations in this paper are done by professional translators. They are all paid at professional rates.

Organizers from the National Research Council Canada and Unbabel have submitted to this task the frozen stable versions of their metrics (YiSi and COMET) dated before this year's shared task and publicly available. Newer versions of COMET were developed without using any of the test set, test suite or challenge sets. We ensured that the metrics co-authored by Tom Kocmi were implemented without using any privileged test sets or insider information.

13 Acknowledgments

Results for this shared task would not be possible without tight collaboration with the organizers of the WMT23 General MT Shared Task. We are grateful to Google and Unbabel for sponsoring and overseeing the human evaluation.

Ricardo Rei is supported by the Portuguese Recovery and Resilience Plan (PRR) through project C645008882-00000055, Center for Responsible AI.

Eleftherios Avramidis is supported by the German Research Foundation (DFG) through the project TextQ (grant num. MO 1038/31-1, 436813723), and by the German Federal Ministry of Education and Research (BMBF) through the project SocialWear (grant num. 01IW2000).

References

Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2023. ACES: Translation accuracy challenge sets at wmt 2023. In *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu,

Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. PaLM 2 Technical Report. *arXiv preprint arXiv:2305.10403*.

Eleftherios Avramidis, Shushen Manakhimova, Vivien Macketanz, and Sebastian Möller. 2023. Challenging the state-of-the-art machine translation metrics from a linguistic perspective. In *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Frédéric Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orăsan, and André F. T. Martins. 2023. Findings of the WMT 2023 Shared Task on Quality Estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Pierre Colombo, Nathan Noiry, Ekhine Irurozki, and Stéphan Clémençon. 2022. What are the best systems? New perspectives on NLP Benchmarking. *arXiv preprint arXiv:2202.03799*.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods. *arXiv preprint arXiv:2104.00054*.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties Matter: Modifying Kendall's Tau for Modern Metric Meta-Evaluation. *arXiv preprint arXiv:2305.14324*.

Sören Dreano, Derek Molloy, and Noel Murphy. 2023a. Embed_Llama: using LLM embeddings for the Metrics Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Sören Dreano, Derek Molloy, and Noel Murphy. 2023b. Tokengram_F, a fast and accurate token-based chrF++ derivative. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. 2017. Replicability analysis for natural language processing: Testing significance with multiple datasets. *Transactions of the Association for Computational Linguistics*, 5:471–486.

Cynthia Dwork, Ravi Kumar, Moni Naor, and D. Sivakumar. 2001. Rank aggregation methods for the web. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, page 613–622, New York, NY, USA. Association for Computing Machinery.

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Bryan Eikema and Wilker Aziz. 2021. Sampling-Based Minimum Bayes Risk Decoding for Neural Machine Translation. *arXiv preprint arXiv:2108.04718*.

Muhammad ElNokrashy and Tom Kocmi. 2023. eBLEU: Unexpectedly Good Machine Translation Evaluation Using Simple Word Embeddings. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Allyson Ettinger, Sudha Rao, Hal Daumé III, and Emily M. Bender. 2017. Towards linguistically generalizable NLP systems: A workshop and shared task. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 1–10, Copenhagen, Denmark. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond English-Centric Multilingual Machine Translation. *Journal of Machine Learning Research*, 22(1).

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2022. MLQE-PE: A multilingual quality estimation and post-editing dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4963–4974, Marseille, France. European Language Resources Association.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. 2023. Epsilon Sampling Rocks: Investigating Sampling Strategies for Minimum Bayes Risk Decoding for Machine Translation. *arXiv preprint arXiv:2305.09860*.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Thamme Gowda, Tom Kocmi, and Marcin Junczys-Dowmunt. 2023. Cometoid: Distilling Strong Reference-based Machine Translation Metrics into Even Stronger Quality Estimation Metrics. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023. xCOMET: Transparent Machine Translation Evaluation through Fine-grained Error Detection. *arXiv preprint arXiv:2310.10482*.

Michael Hagmann and Stefan Riezler. 2023. Towards inferential reproducibility of machine learning research. *arXiv preprint arXiv:2302.04054*.

John Hewitt, Christopher Manning, and Percy Liang. 2022. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3414–3427, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google Submission to the WMT 2023 Metrics Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, Mariya Shmatova, and Jun Suzuki. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022a. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting Translation Quality Error Spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Tom Kocmi, Hitokazu Matsushita, and Christian Federmann. 2022b. MS-COMET: More and better human judgements improve metric performance. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 541–548, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Chi-kiu Lo, Rebecca Knowles, and Cyril Goutte. 2023a. Beyond correlation: Making sense of the score differences of new mt evaluation metrics. In *Proceedings of Machine Translation Summit XIX Vol. 1: Research Track*, pages 186–199.

Chi-kiu Lo, Samuel Larkin, and Rebecca Knowles. 2023b. Metric score landscape challenge (MSLC23): Understanding metrics' performance on a wider landscape of translation quality. In *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional Quality Metrics (MQM) : A

Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica*, pages 0455–463.

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2023. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt. *arXiv preprint arXiv:2303.13809*.

Benjamin Marie. 2022. Yes, we need statistical significance testing. towardsai.net https://pub.towardsai.net/yes-we-need-statistical-significance-testing-927a8d21f9f0.

Ananya Mukherjee and Manish Shrivastava. 2023. MEE4 and XLsim: IIIT HYD's Submissions for WMT23 Metrics Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Mathias Müller and Rico Sennrich. 2021. Understanding the properties of minimum Bayes risk decoding in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 259–272, Online. Association for Computational Linguistics.

Subhajit Naskar, Daniel Deutsch, and Markus Freitag. 2023. Quality Estimation using Minimum Bayes Risk. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv preprint arXiv:2207.04672*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Niccolò Campolungo, and Roberto Navigli. 2022. MaTESe: Machine translation evaluation as a sequence tagging problem. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 569–577, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vikas Raunak, Tom Kocmi, and Matt Post. 2023. SLIDE: Sliding Document Evaluator for Document-Context Evaluation in Machine Translation. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. 2023. Scaling up COMETKIWI: Unbabel-IST 2023 Submission for the Quality Estimation Shared Task. In *Proceedings of the eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

T. Robertson, F.T. Wright, and R. Dykstra. 1988. *Order Restricted Statistical Inference*. Probability and Statistics Series. Wiley.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020a. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020b. Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the Fifth Conference on Machine Translation*, pages 561–570, Online. Association for Computational Linguistics.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Vasiliy Viskov, George Kokush, Daniil Larionov, Steffen Eger, and Alexander Panchenko. 2023. Semantically-Informed Regressive Encoder Score. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Johnny Wei, Tom Kocmi, and Christian Federmann. 2022. Searching for a higher power in the human evaluation of MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 129–139, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhanglin Wu, Yilun Liu, Min Zhang, Xiaofeng Zhao, Junhao Zhu, Ming Zhu, Xiaosong Qiao, Jingfei Zhang, Ma Miaomiao, Zhao Yanqing, Song Peng, shimin tao, Hao Yang, and Yanfei Jiang. 2023. Empowering a Metric with LLM-assisted Named Entity Annotation: HW-TSC's Submission to the WMT23 Metrics Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Jing Xu, Megan Ung, Mojtaba Komeili, Kushal Arora, Y-Lan Boureau, and Jason Weston. 2023a. Learning new skills after deployment: Improving open-domain internet-driven dialogue with human feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13557–13572, Toronto, Canada. Association for Computational Linguistics.

Wenda Xu, Xian Qian, Mingxuan Wang, Lei Li, and William Yang Wang. 2023b. Sescore2: Learning text generation evaluation via synthesizing realistic mistakes.

Wenda Xu, Yilin Tuan, Yujie Lu, Michael Saxon, Lei Li, and William Yang Wang. 2022. Not all errors are equal: Learning text generation metrics using stratified error synthesis. *arXiv preprint arXiv:2210.05035*.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Yang Wang, and Lei Li. 2023c. Instructscore: Explainable text generation evaluation with finegrained feedback.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Correlations with MQM for all metrics

Tables 16 and 17 contain system- and segment-level results for all metrics (including contrastive submissions) on the 10 standard tasks described in Table 7. No pairwise significance tests were carried out for these results, so the per-task ranks only indicate each metric's order on that task, rather than its significance cluster as in Tables 8 and 9.

lang: corr_fcn: metric	avg-corr		en→de,he→en,zh→en accuracy task1		en→de pearson task2		he→en pearson task5		zh→en pearson task8	
XCOMET-Ensemble	**1**	**0.825**	6	0.928	9	0.980	4	0.950	14	0.927
XCOMET-XXL	2	0.824	5	0.932	7	0.982	**1**	**0.964**	16	0.911
*MetricX-23-QE-b**	3	0.823	2	0.940	8	0.982	5	0.947	15	0.926
XCOMET-XL	4	0.816	7	0.924	18	0.973	11	0.937	26	0.884
*MetricX-23-QE-c**	5	0.813	4	0.932	20	0.972	8	0.939	4	0.974
MetricX-23-b	6	0.811	9	0.916	4	0.990	15	0.928	19	0.902
XCOMET-QE-Ensemble*	7	0.808	13	0.908	16	0.974	23	0.909	23	0.892
MetricX-23	8	0.808	12	0.908	12	0.977	22	0.910	28	0.873
GEMBA-MQM*	9	0.802	**1**	**0.944**	**1**	**0.993**	9	0.939	**1**	**0.991**
MetricX-23-QE*	10	0.800	24	0.892	22	0.969	35	0.858	30	0.859
*cometoid22-wmt23**	11	0.794	3	0.936	10	0.979	16	0.928	8	0.956
mbr-metricx-qe*	12	0.788	29	0.880	13	0.976	19	0.915	11	0.936
*CometKiwi-XXL**	13	0.786	11	0.912	6	0.986	14	0.929	2	0.978
*CometKiwi-XL**	14	0.786	8	0.916	14	0.975	29	0.900	3	0.974
MaTESe	15	0.782	17	0.904	36	0.918	25	0.906	25	0.889
CometKiwi*	16	0.782	16	0.904	27	0.946	34	0.860	6	0.963
COMET	17	0.779	20	0.900	3	0.990	7	0.940	21	0.898
MetricX-23-c	18	0.778	10	0.916	28	0.944	6	0.946	9	0.953
instructscore	19	0.777	22	0.896	25	0.952	21	0.910	31	0.825
BLEURT-20	20	0.776	23	0.892	5	0.990	12	0.937	27	0.880
KG-BERTScore*	21	0.774	27	0.884	30	0.926	24	0.908	7	0.962
sescoreX	22	0.772	25	0.892	26	0.952	28	0.901	35	0.797
cometoid22-wmt22*	23	0.772	28	0.880	17	0.973	37	0.839	10	0.940
*cometoid22-wmt21**	24	0.768	30	0.871	19	0.973	38	0.832	13	0.929
docWMT22CometDA	25	0.768	18	0.904	2	0.990	17	0.922	17	0.907
docWMT22CometKiwiDA*	26	0.767	21	0.900	21	0.970	26	0.906	5	0.965
Calibri-COMET22	27	0.767	15	0.904	23	0.963	13	0.930	29	0.863
Calibri-COMET22-QE*	28	0.755	34	0.863	11	0.978	40	0.778	12	0.934
YiSi-1	29	0.754	33	0.871	31	0.925	18	0.917	32	0.823
MS-COMET-QE-22*	30	0.744	32	0.871	24	0.959	43	0.721	20	0.901
prismRef	31	0.744	37	0.851	33	0.920	3	0.956	40	0.762
mre-score-labse-regular	32	0.743	26	0.888	29	0.942	2	0.958	18	0.903
BERTscore	33	0.742	31	0.871	38	0.891	30	0.895	33	0.810
XLsim	34	0.719	36	0.855	32	0.925	31	0.887	36	0.796
f200spBLEU	35	0.704	40	0.819	34	0.919	39	0.805	39	0.772
MEE4	36	0.704	39	0.823	41	0.861	32	0.879	41	0.743
tokengram_F	37	0.703	42	0.815	43	0.858	33	0.878	37	0.795
embed_llama	38	0.701	38	0.831	42	0.861	36	0.841	38	0.785
BLEU	39	0.696	41	0.815	37	0.917	42	0.769	42	0.734
chrF	40	0.694	43	0.795	40	0.866	41	0.776	34	0.809
eBLEU	41	0.692	35	0.859	35	0.918	20	0.911	43	0.727
Random-sysname*	42	0.529	44	0.578	44	0.357	44	0.209	44	0.093
prismSrc*	43	0.455	45	0.386	45	-0.327	45	-0.017	45	-0.406
HuaweiTSC_EE_Metric	–	–	19	0.900	39	0.878	27	0.903	22	0.894
*slide**	–	–	14	0.904	15	0.975	10	0.938	24	0.890

Table 16: Results for all metrics on system-level tasks for main language pairs. Rows are sorted by the overall average correlation across all 10 tasks (leftmost column). Starred metrics are reference-free, underlined metrics are baselines, and italicized metrics are contrastive submissions.

lang: corr_fcn: metric	en→de pearson task3		en→de acc-t task4		he→en pearson task6		he→en acc-t task7		zh→en pearson task9		zh→en acc-t task10	
XCOMET-Ensemble	2	0.695	3	0.604	**1**	**0.556**	**1**	**0.586**	2	0.650	2	0.543
XCOMET-XXL	**1**	**0.695**	4	0.603	2	0.556	4	0.577	5	0.627	3	0.541
*MetricX-23-QE-b**	5	0.628	**1**	**0.606**	6	0.529	3	0.580	**1**	**0.661**	4	0.539
XCOMET-XL	3	0.680	6	0.601	5	0.536	7	0.568	7	0.624	9	0.531
*MetricX-23-QE-c**	11	0.525	11	0.581	7	0.526	6	0.576	9	0.581	**1**	**0.545**
MetricX-23-b	9	0.566	2	0.604	4	0.537	2	0.581	8	0.612	6	0.535
XCOMET-QE-Ensemble*	4	0.679	8	0.588	9	0.498	10	0.554	4	0.647	7	0.533
MetricX-23	7	0.585	5	0.603	3	0.548	5	0.577	6	0.625	8	0.531
GEMBA-MQM*	16	0.502	17	0.572	13	0.401	9	0.564	16	0.449	14	0.522
MetricX-23-QE*	6	0.626	7	0.596	8	0.520	8	0.564	3	0.647	11	0.527
*cometoid22-wmt23**	20	0.448	9	0.586	16	0.397	15	0.544	19	0.439	15	0.520
mbr-metricx-qe*	8	0.571	10	0.584	12	0.411	11	0.553	13	0.489	5	0.537
*CometKiwi-XXL**	28	0.417	13	0.578	19	0.390	13	0.550	24	0.390	10	0.528
*CometKiwi-XL**	21	0.446	18	0.571	22	0.384	18	0.533	21	0.430	13	0.522
MaTESe	10	0.554	30	0.528	10	0.459	12	0.550	11	0.511	34	0.479
CometKiwi*	18	0.475	19	0.569	20	0.387	14	0.544	18	0.442	12	0.525
COMET	25	0.432	15	0.574	14	0.401	19	0.532	22	0.396	19	0.514
MetricX-23-c	15	0.508	27	0.539	31	0.313	20	0.531	27	0.371	21	0.507
instructscore	12	0.519	20	0.563	11	0.458	17	0.536	12	0.499	40	0.459
BLEURT-20	17	0.484	16	0.572	24	0.382	24	0.519	26	0.378	16	0.518
KG-BERTScore*	19	0.451	23	0.556	23	0.382	16	0.537	20	0.430	17	0.516
sescoreX	13	0.519	21	0.563	21	0.385	33	0.484	10	0.536	24	0.499
cometoid22-wmt22*	23	0.441	14	0.578	26	0.365	25	0.515	14	0.479	18	0.515
*cometoid22-wmt21**	26	0.428	12	0.581	27	0.360	26	0.515	15	0.458	20	0.514
docWMT22CometDA	30	0.394	22	0.559	28	0.339	31	0.497	29	0.353	28	0.493
docWMT22CometKiwiDA*	22	0.444	24	0.547	33	0.286	32	0.489	25	0.387	27	0.493
Calibri-COMET22	29	0.413	34	0.522	15	0.401	27	0.515	23	0.396	36	0.474
Calibri-COMET22-QE*	24	0.441	41	0.483	18	0.395	29	0.506	17	0.443	29	0.491
YiSi-1	31	0.366	26	0.542	17	0.395	21	0.529	30	0.290	22	0.504
MS-COMET-QE-22*	33	0.310	25	0.546	32	0.295	30	0.498	28	0.367	26	0.498
prismRef	14	0.516	38	0.518	30	0.319	22	0.528	33	0.183	23	0.504
mre-score-labse-regular	41	0.111	28	0.530	25	0.378	23	0.522	35	0.145	32	0.481
BERTscore	32	0.325	31	0.528	29	0.335	28	0.515	31	0.236	25	0.499
XLsim	35	0.239	32	0.527	35	0.233	34	0.480	37	0.111	39	0.464
f200spBLEU	36	0.237	33	0.526	36	0.230	37	0.447	38	0.108	35	0.476
MEE4	39	0.202	29	0.529	34	0.256	41	0.441	39	0.105	33	0.480
tokengram_F	38	0.227	35	0.520	37	0.226	35	0.461	41	0.060	31	0.485
embed_llama	34	0.250	40	0.483	40	0.215	42	0.430	34	0.161	41	0.447
BLEU	40	0.192	36	0.520	39	0.220	39	0.442	36	0.119	38	0.472
chrF	37	0.232	37	0.519	38	0.221	36	0.460	40	0.063	30	0.485
eBLEU	43	-0.011	39	0.512	42	0.131	38	0.445	43	-0.084	37	0.473
Random-sysname*	42	0.064	43	0.409	43	0.041	43	0.428	42	0.018	43	0.381
prismSrc*	27	0.425	42	0.426	41	0.140	40	0.441	32	0.223	42	0.421

Table 17: Results for all metrics on segment-level tasks for main language pairs. Rows are sorted by the overall average correlation across all 10 tasks (leftmost column in Table 16). Starred metrics are reference-free, underlined metrics are baselines, and italicized metrics are contrastive submissions.

metric		avg corr	p-values																																	
XCOMET-Ensemble	1	0.825	.	01	01	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00
XCOMET-QE-Ensemble*	2	0.808	.	.	46	20	26	00	00	00	01	01	01	03	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00
MetricX-23	2	0.808	.	.	.	24	25	03	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00
GEMBA-MQM*	2	0.802	.	.	.	.	43	03	00	00	00	00	00	02	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00
MetricX-23-QE*	2	0.800	.	.	.	.	.	13	07	05	03	00	06	02	00	00	00	00	01	01	00	00	02	00	00	00	00	00	00	00	00	00	00	00	00	00
mbr-metricx-qe*	3	0.788	.	.	.	.	.	.	31	24	17	10	16	09	02	02	00	00	00	03	00	02	01	00	00	00	00	00	00	00	00	00	00	00	00	00
MaTESe	3	0.782	.	.	.	.	.	.	.	48	38	26	19	24	12	09	04	06	03	03	00	01	03	00	00	00	00	00	00	00	00	00	00	00	00	00
CometKiwi*	3	0.782	.	.	.	.	.	.	.	.	39	25	26	23	04	07	01	02	02	02	00	00	01	00	00	00	00	00	00	00	00	00	00	00	00	00	00
COMET	3	0.779	.	.	.	.	.	.	.	.	.	22	34	25	23	01	19	11	11	01	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00
BLEURT-20	3	0.776	.	.	.	.	.	.	.	.	.	.	46	34	35	10	20	16	13	04	02	00	00	01	00	00	00	00	00	00	00	00	00	00	00	00	00
KG-BERTScore*	3	0.774	.	.	.	.	.	.	.	.	.	.	.	43	49	24	29	32	16	08	00	04	07	02	00	00	00	00	00	00	00	00	00	00	00	00	00
sescoreX	3	0.772	.	.	.	.	.	.	.	.	.	.	.	.	49	30	37	31	18	06	08	03	04	04	00	00	00	00	00	00	00	00	00	00	00	00	00
cometoid22-wmt22*	4	0.772	.	.	.	.	.	.	.	.	.	.	.	.	.	34	22	22	07	14	03	04	07	01	00	00	00	00	00	00	00	00	00	00	00	00	00
docWMT22CometDA	4	0.768	.	.	.	.	.	.	.	.	.	.	.	.	.	.	51	44	24	10	03	03	03	04	00	00	00	00	00	00	00	00	00	00	00	00	00
docWMT22CometKiwiDA*	4	0.767	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	48	14	20	07	09	12	03	00	00	00	00	00	00	00	00	00	00	00	00	00
Calibri-COMET22	4	0.767	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	17	23	10	16	11	01	00	00	00	00	00	00	00	00	00	00	00	00	00
Calibri-COMET22-QE*	4	0.755	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	45	30	36	30	18	07	01	04	02	01	00	00	00	00	00	00	00	00
YiSi-1	4	0.754	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	30	13	22	31	00	00	00	00	00	00	00	00	00	00	00	00	00
MS-COMET-QE-22*	5	0.744	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	52	49	43	12	01	02	02	02	01	00	00	00	00	00	00	00
prismRef	5	0.744	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	44	44	00	00	01	00	00	00	00	00	00	00	00	00	00
mre-score-labse-regular	5	0.743	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	49	06	01	04	01	00	00	00	00	00	00	00	00	00
BERTscore	5	0.742	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	18	03	07	05	01	02	02	00	00	00	00	00	00
XLsim	6	0.719	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	04	10	01	06	01	01	00	00	00	00	00	00
f200spBLEU	7	0.704	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	51	48	39	06	13	12	00	00	00	00	00
MEE4	7	0.704	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	46	46	33	23	16	00	00	00	00	00
tokengram_F	7	0.703	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	45	22	15	11	00	00	00	00	00
embed_llama	7	0.701	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	34	30	29	00	00	00	00	00
BLEU	7	0.696	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	42	35	00	00	00	00	00
chrF	7	0.694	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	43	00	00	00	00	00
eBLEU	7	0.692	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	00	00	00	00	00
Random-sysname*	8	0.529	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	04
prismSrc*	9	0.455	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

Table 18: Results of pairwise metric significance tests for primary submissions using permutation resampling. Each value gives the 100 $\times$ estimated probability of the null hypothesis that the average correlation of the metric in the current row is $\leq$ the average correlation of the metric in the current column. Starred metrics are reference-free, and underlined metrics are baselines.

B Significance comparisons for main results

Table 18 contains the results of pairwise comparisons for the results in Table 1.

C Correlations with WMT DA-SQM for all metrics

Tables 19 to 27 give correlations with WMT direct assessment (DA-SQM) scores on all 8 translation directions for which those scores are available. In all cases, reference A was used, and no additional metrics were available to be scored by the metrics. We evaluate metrics on a task setup similar to that of Table 7: one system-level pairwise accuracy task involving all languages (with a weight of 8), and system-level Pearson, segment-level Pearson, and segment-level acc^*_{eq} tasks for each translation direction (24 tasks in total, each with a weight of 1). Each table shows overall average correlation, along with the results for the tasks for one translation direction. Metrics that did not participate in all tasks do not have an average correlation, and are displayed at the end of each table.

We wish to emphasize that the DA+SQM is considerably noisier than MQM. This increased variability may influence the outcomes observed in the following spotlight evaluation. Consequently, readers should exercise considerable caution when drawing conclusions from these results.

lang: corr_fcn: metric	avg-corr		cs→uk,de→en,en→cs,en→de,en→ja,en→zh,ja→en,zh→en accuracy task1	
*CometKiwi-XXL**	**1**	**0.798**	**1**	**0.912**
*CometKiwi-XL**	2	0.795	2	0.905
COMET	3	0.787	8	0.890
CometKiwi*	4	0.787	10	0.887
*cometoid22-wmt23**	5	0.786	6	0.897
KG-BERTScore*	6	0.784	12	0.884
*MetricX-23-QE-c**	7	0.780	9	0.887
BLEURT-20	8	0.778	15	0.880
*MetricX-23-QE-b**	9	0.777	14	0.880
cometoid22-wmt22*	10	0.776	13	0.884
MetricX-23-c	11	0.775	5	0.898
*cometoid22-wmt21**	12	0.774	11	0.885
XCOMET-Ensemble	13	0.774	20	0.870
MetricX-23-b	14	0.768	17	0.873
MetricX-23-QE*	15	0.768	19	0.870
MS-COMET-QE-22*	16	0.767	16	0.879
XCOMET-QE-Ensemble*	17	0.766	18	0.871
MetricX-23	18	0.762	22	0.863
YiSi-1	19	0.749	25	0.832
XCOMET-XL	20	0.748	24	0.860
XLsim	21	0.745	26	0.831
XCOMET-XXL	22	0.743	21	0.866
GEMBA-MQM*	23	0.739	4	0.899
prismRef	24	0.736	27	0.808
mre-score-labse-regular	25	0.734	23	0.861
BERTscore	26	0.732	28	0.799
tokengram_F	27	0.714	30	0.786
chrF	28	0.712	33	0.777
f200spBLEU	29	0.708	29	0.786
embed_llama	30	0.701	32	0.778
eBLEU	31	0.694	31	0.781
BLEU	32	0.660	34	0.770
Random-sysname*	33	0.537	35	0.580
prismSrc*	34	0.514	36	0.412
HuaweiTSC_EE_Metric	–	–	7	0.892
*slide**	–	–	3	0.902

Table 19: Correlations with WMT DA-SQM scores for all metrics on all-pairs data. Rows are sorted by the overall average correlation across all 25 tasks (leftmost column). Starred metrics are reference-free, underlined metrics are baselines, and italicized metrics are contrastive submissions.

lang: corr_fcn: metric	avg-corr		cs→uk pearson task1		cs→uk pearson task2		cs→uk acc-t task3	
*CometKiwi-XXL**	**1**	**0.798**	12	0.889	4	0.462	6	0.555
*CometKiwi-XL**	2	0.795	18	0.866	15	0.412	10	0.548
COMET	3	0.787	5	0.899	6	0.454	8	0.553
CometKiwi*	4	0.787	23	0.788	8	0.429	13	0.536
*cometoid22-wmt23**	5	0.786	7	0.898	11	0.420	15	0.534
KG-BERTScore*	6	0.784	24	0.788	9	0.429	16	0.530
*MetricX-23-QE-c**	7	0.780	3	0.920	2	0.502	9	0.553
BLEURT-20	8	0.778	2	0.926	7	0.443	12	0.538
*MetricX-23-QE-b**	9	0.777	6	0.898	16	0.410	4	0.559
cometoid22-wmt22*	10	0.776	19	0.851	19	0.403	19	0.528
MetricX-23-c	11	0.775	**1**	**0.932**	**1**	**0.523**	5	0.558
*cometoid22-wmt21**	12	0.774	21	0.822	14	0.414	23	0.521
XCOMET-Ensemble	13	0.774	8	0.897	3	0.482	3	0.560
MetricX-23-b	14	0.768	13	0.888	17	0.410	**1**	**0.568**
MetricX-23-QE*	15	0.768	11	0.889	21	0.382	7	0.555
MS-COMET-QE-22*	16	0.767	20	0.851	23	0.322	24	0.519
XCOMET-QE-Ensemble*	17	0.766	17	0.873	5	0.462	11	0.540
MetricX-23	18	0.762	15	0.879	20	0.395	2	0.567
YiSi-1	19	0.749	26	0.753	25	0.315	20	0.526
XCOMET-XL	20	0.748	14	0.882	10	0.423	18	0.529
XLsim	21	0.745	22	0.792	24	0.318	21	0.526
XCOMET-XXL	22	0.743	9	0.897	18	0.407	33	0.436
GEMBA-MQM*	23	0.739	4	0.913	12	0.419	34	0.323
prismRef	24	0.736	27	0.694	22	0.372	17	0.530
mre-score-labse-regular	25	0.734	25	0.772	13	0.417	14	0.534
BERTscore	26	0.732	32	0.544	26	0.292	22	0.524
tokengram_F	27	0.714	30	0.626	28	0.268	25	0.518
chrF	28	0.712	29	0.637	27	0.273	26	0.517
f200spBLEU	29	0.708	28	0.676	30	0.221	28	0.504
embed_llama	30	0.701	34	0.511	33	0.157	30	0.492
eBLEU	31	0.694	33	0.512	31	0.188	27	0.511
BLEU	32	0.660	31	0.548	32	0.184	31	0.480
Random-sysname*	33	0.537	35	0.343	34	0.047	32	0.469
prismSrc*	34	0.514	36	-0.236	29	0.261	29	0.495
HuaweiTSC_EE_Metric	–	–	10	0.893	–	–	–	–
*slide**	–	–	16	0.877	–	–	–	–

Table 20: Correlations with WMT DA-SQM scores for all metrics on cs→uk data. Rows are sorted by the overall average correlation across all 25 tasks (leftmost column). Starred metrics are reference-free, underlined metrics are baselines, and italicized metrics are contrastive submissions.

lang: corr_fcn: metric	avg-corr		de→en pearson task1		de→en pearson task2		de→en acc-t task3	
*CometKiwi-XXL**	**1**	**0.798**	15	0.931	14	0.411	11	0.571
*CometKiwi-XL**	2	0.795	12	0.934	17	0.402	13	0.569
COMET	3	0.787	6	0.953	2	0.480	2	0.584
CometKiwi*	4	0.787	14	0.933	9	0.447	16	0.559
*cometoid22-wmt23**	5	0.786	17	0.913	3	0.471	12	0.571
KG-BERTScore*	6	0.784	13	0.933	7	0.447	20	0.553
*MetricX-23-QE-c**	7	0.780	30	0.835	8	0.447	7	0.574
BLEURT-20	8	0.778	3	0.965	**1**	**0.486**	5	0.578
*MetricX-23-QE-b**	9	0.777	21	0.893	12	0.425	4	0.579
cometoid22-wmt22*	10	0.776	23	0.881	6	0.449	17	0.558
MetricX-23-c	11	0.775	9	0.944	27	0.298	24	0.544
*cometoid22-wmt21**	12	0.774	26	0.856	10	0.437	19	0.556
XCOMET-Ensemble	13	0.774	28	0.842	15	0.408	9	0.573
MetricX-23-b	14	0.768	27	0.850	18	0.389	**1**	**0.590**
MetricX-23-QE*	15	0.768	24	0.876	13	0.418	8	0.574
MS-COMET-QE-22*	16	0.767	29	0.841	32	0.256	23	0.545
XCOMET-QE-Ensemble*	17	0.766	34	0.813	19	0.385	18	0.556
MetricX-23	18	0.762	31	0.831	20	0.382	3	0.584
YiSi-1	19	0.749	**1**	**0.970**	5	0.451	10	0.572
XCOMET-XL	20	0.748	35	0.780	22	0.341	25	0.544
XLsim	21	0.745	8	0.947	23	0.340	15	0.560
XCOMET-XXL	22	0.743	32	0.828	21	0.375	31	0.517
GEMBA-MQM*	23	0.739	10	0.938	4	0.463	34	0.426
prismRef	24	0.736	4	0.963	16	0.403	14	0.565
mre-score-labse-regular	25	0.734	16	0.916	34	0.121	26	0.540
BERTscore	26	0.732	2	0.969	11	0.434	6	0.576
tokengram_F	27	0.714	22	0.891	25	0.319	21	0.551
chrF	28	0.712	25	0.860	24	0.328	22	0.550
f200spBLEU	29	0.708	19	0.904	28	0.291	27	0.539
embed_llama	30	0.701	18	0.913	29	0.275	30	0.525
eBLEU	31	0.694	5	0.954	33	0.207	28	0.538
BLEU	32	0.660	20	0.897	31	0.270	29	0.534
Random-sysname*	33	0.537	37	0.185	35	0.044	33	0.472
prismSrc*	34	0.514	36	0.449	30	0.273	32	0.502
HuaweiTSC_EE_Metric	–	–	7	0.950	–	–	–	–
*slide**	–	–	11	0.934	–	–	–	–
MaTESe	–	–	33	0.816	26	0.308	35	0.373

Table 21: Correlations with WMT DA-SQM scores for all metrics on de→en data. Rows are sorted by the overall average correlation across all 25 tasks (leftmost column). Starred metrics are reference-free, underlined metrics are baselines, and italicized metrics are contrastive submissions.

lang: corr_fcn: metric	avg-corr		en→cs pearson task1		en→cs pearson task2		en→cs acc-t task3	
*CometKiwi-XXL**	**1**	**0.798**	**1**	**0.922**	7	0.367	5	0.548
*CometKiwi-XL**	2	0.795	5	0.897	6	0.369	7	0.541
COMET	3	0.787	14	0.865	4	0.377	10	0.524
CometKiwi*	4	0.787	22	0.790	13	0.350	11	0.518
*cometoid22-wmt23**	5	0.786	13	0.865	11	0.352	16	0.507
KG-BERTScore*	6	0.784	21	0.790	12	0.350	17	0.507
*MetricX-23-QE-c**	7	0.780	6	0.893	3	0.391	8	0.540
BLEURT-20	8	0.778	20	0.793	5	0.373	12	0.510
*MetricX-23-QE-b**	9	0.777	10	0.881	18	0.338	2	0.551
cometoid22-wmt22*	10	0.776	17	0.825	16	0.341	18	0.506
MetricX-23-c	11	0.775	23	0.750	19	0.316	13	0.510
*cometoid22-wmt21**	12	0.774	18	0.824	17	0.340	14	0.508
XCOMET-Ensemble	13	0.774	3	0.903	**1**	**0.402**	6	0.543
MetricX-23-b	14	0.768	11	0.880	15	0.344	**1**	**0.552**
MetricX-23-QE*	15	0.768	12	0.878	14	0.348	4	0.549
MS-COMET-QE-22*	16	0.767	19	0.797	21	0.286	21	0.497
XCOMET-QE-Ensemble*	17	0.766	2	0.908	2	0.395	9	0.528
MetricX-23	18	0.762	7	0.891	9	0.361	3	0.550
YiSi-1	19	0.749	26	0.568	24	0.245	24	0.492
XCOMET-XL	20	0.748	4	0.898	8	0.362	15	0.507
XLsim	21	0.745	25	0.627	23	0.259	20	0.503
XCOMET-XXL	22	0.743	8	0.890	10	0.353	33	0.439
GEMBA-MQM*	23	0.739	16	0.852	20	0.309	34	0.327
prismRef	24	0.736	27	0.557	22	0.265	22	0.495
mre-score-labse-regular	25	0.734	24	0.718	33	0.130	19	0.504
BERTscore	26	0.732	30	0.480	25	0.228	23	0.493
tokengram_F	27	0.714	34	0.409	26	0.203	26	0.481
chrF	28	0.712	33	0.450	27	0.201	27	0.480
f200spBLEU	29	0.708	29	0.496	28	0.199	29	0.475
embed_llama	30	0.701	32	0.466	30	0.172	28	0.476
eBLEU	31	0.694	31	0.467	32	0.169	25	0.483
BLEU	32	0.660	28	0.519	29	0.186	30	0.460
Random-sysname*	33	0.537	35	0.015	34	0.002	32	0.452
prismSrc*	34	0.514	36	-0.042	31	0.171	31	0.456
HuaweiTSC_EE_Metric	–	–	15	0.862	–	–	–	–
*slide**	–	–	9	0.885	–	–	–	–

Table 22: Correlations with WMT DA-SQM scores for all metrics on en→cs data. Rows are sorted by the overall average correlation across all 25 tasks (leftmost column). Starred metrics are reference-free, underlined metrics are baselines, and italicized metrics are contrastive submissions.

lang: corr_fcn: metric	avg-corr		en→de pearson task1		en→de pearson task2		en→de acc-t task3	
*CometKiwi-XXL**	**1**	**0.798**	13	0.972	4	0.506	**1**	**0.595**
*CometKiwi-XL**	2	0.795	5	0.984	3	0.512	3	0.589
COMET	3	0.787	21	0.953	5	0.496	5	0.588
CometKiwi*	4	0.787	**1**	**0.990**	**1**	**0.537**	6	0.586
*cometoid22-wmt23**	5	0.786	26	0.944	6	0.491	12	0.580
KG-BERTScore*	6	0.784	3	0.990	2	0.523	17	0.578
*MetricX-23-QE-c**	7	0.780	39	0.859	12	0.465	19	0.576
BLEURT-20	8	0.778	25	0.945	15	0.452	18	0.577
*MetricX-23-QE-b**	9	0.777	33	0.910	19	0.437	4	0.588
cometoid22-wmt22*	10	0.776	32	0.911	16	0.447	21	0.575
MetricX-23-c	11	0.775	2	0.990	8	0.482	13	0.580
*cometoid22-wmt21**	12	0.774	34	0.905	20	0.433	22	0.574
XCOMET-Ensemble	13	0.774	38	0.861	25	0.399	20	0.576
MetricX-23-b	14	0.768	35	0.896	30	0.377	8	0.583
MetricX-23-QE*	15	0.768	37	0.867	18	0.443	11	0.582
MS-COMET-QE-22*	16	0.767	28	0.942	33	0.371	28	0.558
XCOMET-QE-Ensemble*	17	0.766	41	0.849	29	0.382	27	0.564
MetricX-23	18	0.762	40	0.855	28	0.389	9	0.582
YiSi-1	19	0.749	6	0.980	13	0.456	23	0.571
XCOMET-XL	20	0.748	42	0.845	34	0.365	32	0.552
XLsim	21	0.745	7	0.979	27	0.391	25	0.566
XCOMET-XXL	22	0.743	36	0.868	24	0.399	39	0.525
GEMBA-MQM*	23	0.739	17	0.961	7	0.488	42	0.434
prismRef	24	0.736	16	0.963	37	0.321	36	0.544
mre-score-labse-regular	25	0.734	30	0.927	42	0.144	35	0.548
BERTscore	26	0.732	12	0.973	23	0.417	24	0.567
tokengram_F	27	0.714	27	0.943	32	0.371	30	0.556
chrF	28	0.712	24	0.945	31	0.374	31	0.553
f200spBLEU	29	0.708	14	0.970	36	0.324	29	0.557
embed_llama	30	0.701	23	0.951	35	0.348	34	0.550
eBLEU	31	0.694	31	0.920	41	0.159	37	0.542
BLEU	32	0.660	18	0.958	38	0.275	38	0.541
Random-sysname*	33	0.537	44	0.278	43	0.075	41	0.482
prismSrc*	34	0.514	45	-0.364	40	0.190	40	0.485
HuaweiTSC_EE_Metric	–	–	10	0.975	–	–	–	–
instructscore	–	–	8	0.977	10	0.473	15	0.578
*slide**	–	–	4	0.984	–	–	–	–
Calibri-COMET22	–	–	22	0.953	21	0.425	7	0.584
Calibri-COMET22-QE*	–	–	19	0.957	17	0.445	33	0.551
MEE4	–	–	15	0.968	22	0.421	26	0.565
MaTESe	–	–	43	0.791	39	0.272	43	0.375
docWMT22CometDA	–	–	29	0.941	14	0.454	2	0.593
docWMT22CometKiwiDA*	–	–	11	0.973	26	0.392	14	0.579
mbr-metricx-qe*	–	–	20	0.954	9	0.477	10	0.582
sescoreX	–	–	9	0.977	11	0.473	16	0.578

Table 23: Correlations with WMT DA-SQM scores for all metrics on en→de data. Rows are sorted by the overall average correlation across all 25 tasks (leftmost column). Starred metrics are reference-free, underlined metrics are baselines, and italicized metrics are contrastive submissions.

lang: corr_fcn: metric	avg-corr		en→ja pearson task1		en→ja pearson task2		en→ja acc-t task3	
*CometKiwi-XXL**	**1**	**0.798**	3	0.993	2	0.527	2	0.592
*CometKiwi-XL**	2	0.795	2	0.993	**1**	**0.528**	**1**	**0.593**
COMET	3	0.787	12	0.969	6	0.462	11	0.580
CometKiwi*	4	0.787	6	0.984	4	0.516	4	0.588
*cometoid22-wmt23**	5	0.786	11	0.979	10	0.449	13	0.574
KG-BERTScore*	6	0.784	5	0.984	3	0.516	7	0.583
*MetricX-23-QE-c**	7	0.780	22	0.955	8	0.456	9	0.580
BLEURT-20	8	0.778	4	0.990	15	0.417	15	0.569
*MetricX-23-QE-b**	9	0.777	21	0.956	14	0.428	3	0.590
cometoid22-wmt22*	10	0.776	20	0.960	11	0.449	14	0.569
MetricX-23-c	11	0.775	28	0.918	23	0.371	25	0.545
*cometoid22-wmt21**	12	0.774	16	0.964	12	0.442	16	0.568
XCOMET-Ensemble	13	0.774	26	0.920	5	0.470	6	0.586
MetricX-23-b	14	0.768	23	0.941	16	0.413	5	0.587
MetricX-23-QE*	15	0.768	30	0.898	17	0.411	10	0.580
MS-COMET-QE-22*	16	0.767	9	0.983	7	0.458	18	0.565
XCOMET-QE-Ensemble*	17	0.766	31	0.895	9	0.455	12	0.574
MetricX-23	18	0.762	29	0.916	18	0.401	8	0.580
YiSi-1	19	0.749	7	0.984	21	0.382	20	0.561
XCOMET-XL	20	0.748	34	0.821	19	0.397	21	0.558
XLsim	21	0.745	27	0.918	24	0.354	22	0.557
XCOMET-XXL	22	0.743	32	0.871	20	0.394	31	0.485
GEMBA-MQM*	23	0.739	8	0.983	13	0.429	33	0.389
prismRef	24	0.736	25	0.922	22	0.371	19	0.561
mre-score-labse-regular	25	0.734	10	0.979	31	0.120	17	0.566
BERTscore	26	0.732	18	0.962	26	0.317	23	0.550
tokengram_F	27	0.714	13	0.969	27	0.227	24	0.548
chrF	28	0.712	14	0.966	28	0.220	26	0.543
f200spBLEU	29	0.708	19	0.961	30	0.190	29	0.523
embed_llama	30	0.701	15	0.964	29	0.212	28	0.524
eBLEU	31	0.694	24	0.926	32	0.073	30	0.522
BLEU	32	0.660	33	0.833	34	0.001	34	0.070
Random-sysname*	33	0.537	36	0.307	33	0.064	32	0.484
prismSrc*	34	0.514	35	0.764	25	0.322	27	0.530
HuaweiTSC_EE_Metric	–	–	17	0.963	–	–	–	–
*slide**	–	–	**1**	**0.995**	–	–	–	–

Table 24: Correlations with WMT DA-SQM scores for all metrics on en→ja data. Rows are sorted by the overall average correlation across all 25 tasks (leftmost column). Starred metrics are reference-free, underlined metrics are baselines, and italicized metrics are contrastive submissions.

lang: corr_fcn: metric	avg-corr		en→zh pearson task1		en→zh pearson task2		en→zh acc-t task3	
*CometKiwi-XXL**	**1**	**0.798**	14	0.982	7	0.559	3	0.601
*CometKiwi-XL**	2	0.795	8	0.988	4	0.588	**1**	**0.601**
COMET	3	0.787	3	0.995	6	0.575	8	0.589
CometKiwi*	4	0.787	4	0.994	**1**	**0.635**	7	0.590
*cometoid22-wmt23**	5	0.786	**1**	**0.997**	5	0.588	11	0.584
KG-BERTScore*	6	0.784	5	0.994	2	0.635	10	0.584
*MetricX-23-QE-c**	7	0.780	28	0.913	18	0.468	12	0.582
BLEURT-20	8	0.778	9	0.988	8	0.550	18	0.571
*MetricX-23-QE-b**	9	0.777	19	0.963	19	0.456	2	0.601
cometoid22-wmt22*	10	0.776	7	0.989	9	0.537	15	0.574
MetricX-23-c	11	0.775	24	0.937	12	0.507	23	0.563
*cometoid22-wmt21**	12	0.774	10	0.988	10	0.527	16	0.573
XCOMET-Ensemble	13	0.774	21	0.944	14	0.493	4	0.596
MetricX-23-b	14	0.768	27	0.926	23	0.420	5	0.595
MetricX-23-QE*	15	0.768	22	0.943	22	0.439	6	0.594
MS-COMET-QE-22*	16	0.767	2	0.996	3	0.610	19	0.570
XCOMET-QE-Ensemble*	17	0.766	30	0.908	21	0.450	14	0.577
MetricX-23	18	0.762	33	0.885	24	0.411	9	0.588
YiSi-1	19	0.749	15	0.977	15	0.493	20	0.566
XCOMET-XL	20	0.748	35	0.790	26	0.366	28	0.542
XLsim	21	0.745	12	0.985	11	0.524	17	0.572
XCOMET-XXL	22	0.743	32	0.885	25	0.391	31	0.517
GEMBA-MQM*	23	0.739	18	0.973	16	0.489	33	0.385
prismRef	24	0.736	17	0.975	13	0.496	21	0.564
mre-score-labse-regular	25	0.734	11	0.986	32	0.177	13	0.577
BERTscore	26	0.732	16	0.975	17	0.474	22	0.563
tokengram_F	27	0.714	20	0.945	27	0.343	24	0.558
chrF	28	0.712	25	0.934	29	0.326	25	0.550
f200spBLEU	29	0.708	31	0.905	28	0.327	26	0.547
embed_llama	30	0.701	26	0.927	30	0.297	27	0.542
eBLEU	31	0.694	29	0.912	31	0.210	29	0.535
BLEU	32	0.660	34	0.804	33	0.093	34	0.141
Random-sysname*	33	0.537	36	0.046	34	0.018	32	0.462
prismSrc*	34	0.514	23	0.941	20	0.452	30	0.527
HuaweiTSC_EE_Metric	–	–	6	0.992	–	–	–	–
*slide**	–	–	13	0.982	–	–	–	–

Table 25: Correlations with WMT DA-SQM scores for all metrics on en→zh data. Rows are sorted by the overall average correlation across all 25 tasks (leftmost column). Starred metrics are reference-free, underlined metrics are baselines, and italicized metrics are contrastive submissions.

lang: corr_fcn: metric	avg-corr		ja→en pearson task1		ja→en pearson task2		ja→en acc-t task3	
*CometKiwi-XXL**	**1**	**0.798**	2	0.984	**1**	**0.474**	2	0.578
*CometKiwi-XL**	2	0.795	3	0.982	4	0.446	6	0.573
COMET	3	0.787	16	0.968	5	0.445	3	0.576
CometKiwi*	4	0.787	10	0.975	3	0.455	10	0.568
*cometoid22-wmt23**	5	0.786	19	0.966	7	0.435	15	0.560
KG-BERTScore*	6	0.784	9	0.975	2	0.455	12	0.561
*MetricX-23-QE-c**	7	0.780	21	0.965	11	0.418	7	0.572
BLEURT-20	8	0.778	22	0.964	6	0.436	9	0.570
*MetricX-23-QE-b**	9	0.777	12	0.972	16	0.383	4	0.575
cometoid22-wmt22*	10	0.776	25	0.946	8	0.432	19	0.550
MetricX-23-c	11	0.775	11	0.972	22	0.342	24	0.547
*cometoid22-wmt21**	12	0.774	26	0.944	9	0.431	20	0.549
XCOMET-Ensemble	13	0.774	24	0.947	12	0.410	5	0.574
MetricX-23-b	14	0.768	29	0.938	21	0.343	**1**	**0.578**
MetricX-23-QE*	15	0.768	30	0.936	20	0.344	11	0.567
MS-COMET-QE-22*	16	0.767	34	0.916	14	0.388	22	0.548
XCOMET-QE-Ensemble*	17	0.766	31	0.935	13	0.388	16	0.557
MetricX-23	18	0.762	33	0.918	24	0.332	8	0.572
YiSi-1	19	0.749	6	0.978	15	0.383	13	0.561
XCOMET-XL	20	0.748	32	0.922	25	0.327	23	0.547
XLsim	21	0.745	**1**	**0.989**	23	0.342	18	0.552
XCOMET-XXL	22	0.743	27	0.941	18	0.352	31	0.492
GEMBA-MQM*	23	0.739	4	0.982	10	0.421	34	0.395
prismRef	24	0.736	13	0.971	19	0.351	17	0.557
mre-score-labse-regular	25	0.734	5	0.980	33	0.186	21	0.548
BERTscore	26	0.732	7	0.977	17	0.357	14	0.560
tokengram_F	27	0.714	18	0.967	27	0.290	25	0.546
chrF	28	0.712	20	0.966	26	0.292	26	0.545
f200spBLEU	29	0.708	23	0.955	29	0.226	28	0.528
embed_llama	30	0.701	14	0.969	31	0.203	29	0.524
eBLEU	31	0.694	15	0.969	32	0.202	27	0.530
BLEU	32	0.660	28	0.939	30	0.221	30	0.517
Random-sysname*	33	0.537	36	0.288	35	0.061	32	0.481
prismSrc*	34	0.514	37	-0.747	34	0.171	33	0.470
HuaweiTSC_EE_Metric	–	–	17	0.967	–	–	–	–
*slide**	–	–	8	0.976	–	–	–	–
MaTESe	–	–	35	0.904	28	0.242	35	0.326

Table 26: Correlations with WMT DA-SQM scores for all metrics on ja→en data. Rows are sorted by the overall average correlation across all 25 tasks (leftmost column). Starred metrics are reference-free, underlined metrics are baselines, and italicized metrics are contrastive submissions.

lang: corr_fcn: metric	avg-corr		zh→en pearson task1		zh→en pearson task2		zh→en acc-t task3	
*CometKiwi-XXL**	**1**	**0.798**	**1**	**0.938**	3	0.435	4	0.540
*CometKiwi-XL**	2	0.795	3	0.936	6	0.427	9	0.535
COMET	3	0.787	21	0.811	11	0.378	15	0.525
CometKiwi*	4	0.787	4	0.931	**1**	**0.460**	10	0.534
*cometoid22-wmt23**	5	0.786	9	0.913	10	0.402	16	0.523
KG-BERTScore*	6	0.784	5	0.927	2	0.448	18	0.521
*MetricX-23-QE-c**	7	0.780	14	0.843	13	0.373	6	0.537
BLEURT-20	8	0.778	25	0.766	17	0.331	21	0.520
*MetricX-23-QE-b**	9	0.777	17	0.823	21	0.298	**1**	**0.544**
cometoid22-wmt22*	10	0.776	8	0.918	5	0.432	19	0.520
MetricX-23-c	11	0.775	7	0.924	16	0.339	24	0.512
*cometoid22-wmt21**	12	0.774	10	0.908	7	0.419	20	0.520
XCOMET-Ensemble	13	0.774	20	0.816	18	0.322	7	0.537
MetricX-23-b	14	0.768	26	0.759	26	0.261	3	0.540
MetricX-23-QE*	15	0.768	24	0.770	22	0.284	5	0.538
MS-COMET-QE-22*	16	0.767	6	0.927	8	0.418	22	0.519
XCOMET-QE-Ensemble*	17	0.766	22	0.803	19	0.315	17	0.522
MetricX-23	18	0.762	30	0.735	24	0.264	8	0.536
YiSi-1	19	0.749	31	0.715	25	0.263	28	0.511
XCOMET-XL	20	0.748	28	0.758	27	0.254	27	0.512
XLsim	21	0.745	32	0.702	33	0.218	26	0.512
XCOMET-XXL	22	0.743	23	0.787	23	0.275	39	0.463
GEMBA-MQM*	23	0.739	11	0.873	14	0.370	41	0.356
prismRef	24	0.736	41	0.632	31	0.229	25	0.512
mre-score-labse-regular	25	0.734	18	0.817	38	0.146	30	0.509
BERTscore	26	0.732	33	0.702	30	0.236	23	0.515
tokengram_F	27	0.714	37	0.670	37	0.167	31	0.503
chrF	28	0.712	35	0.701	35	0.168	32	0.503
f200spBLEU	29	0.708	39	0.651	39	0.139	36	0.483
embed_llama	30	0.701	34	0.702	41	0.123	34	0.494
eBLEU	31	0.694	42	0.629	42	0.107	35	0.494
BLEU	32	0.660	43	0.610	40	0.134	37	0.475
Random-sysname*	33	0.537	44	-0.144	43	-0.026	40	0.446
prismSrc*	34	0.514	45	-0.457	28	0.248	38	0.471
HuaweiTSC_EE_Metric	–	–	19	0.816	–	–	–	–
instructscore	–	–	38	0.652	32	0.227	42	0.342
*slide**	–	–	12	0.863	–	–	–	–
Calibri-COMET22	–	–	27	0.759	20	0.313	13	0.529
Calibri-COMET22-QE*	–	–	13	0.854	12	0.375	11	0.530
MEE4	–	–	40	0.632	36	0.168	33	0.498
MaTESe	–	–	29	0.739	34	0.201	43	0.319
docWMT22CometDA	–	–	15	0.836	15	0.345	12	0.530
docWMT22CometKiwiDA*	–	–	2	0.938	9	0.403	2	0.542
mbr-metricx-qe*	–	–	16	0.827	4	0.435	14	0.526
sescoreX	–	–	36	0.695	29	0.238	29	0.509

Table 27: Correlations with WMT DA-SQM scores for all metrics on zh→en data. Rows are sorted by the overall average correlation across all 25 tasks (leftmost column). Starred metrics are reference-free, underlined metrics are baselines, and italicized metrics are contrastive submissions.

D Additional figures

Figures 9-14 show the (log) p-value of one-sided paired t-test on the MQM scores against the score difference of each metric for each system pair in each translation direction. Figures 15-20 show the (log) p-value of significance test with bootstrap resampling on the metric scores against the score difference of that metric for each system pair in each translation direction.

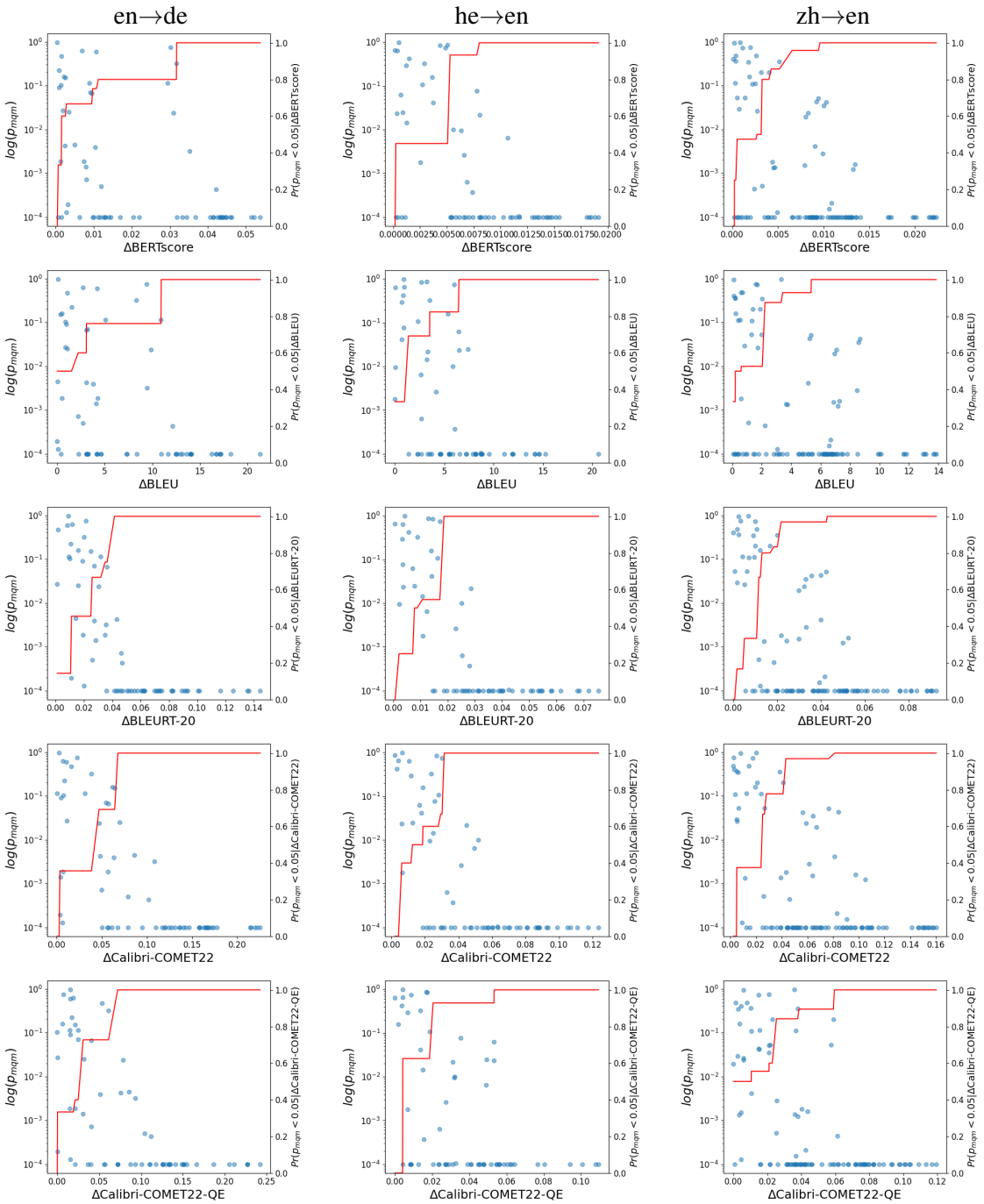


Figure 9: Log p-value of one-sided paired t-test on MQM scores (p_{mqm}) against the score difference of each metric (top to bottom: BERTScore, BLEU, BLEURT-20, CALIBRI-COMET22, CALIBRI-COMET22-QE) for each system pair in each translation direction (left to right: en→de, he→en, zh→en). The red line is the isotonic regression fit to all data points, representing $Pr(p_{mqm} < 0.05|\Delta M)$. Note: for readability, values of p_{mqm} are rounded up to 0.0001 when they are less than 0.0001.

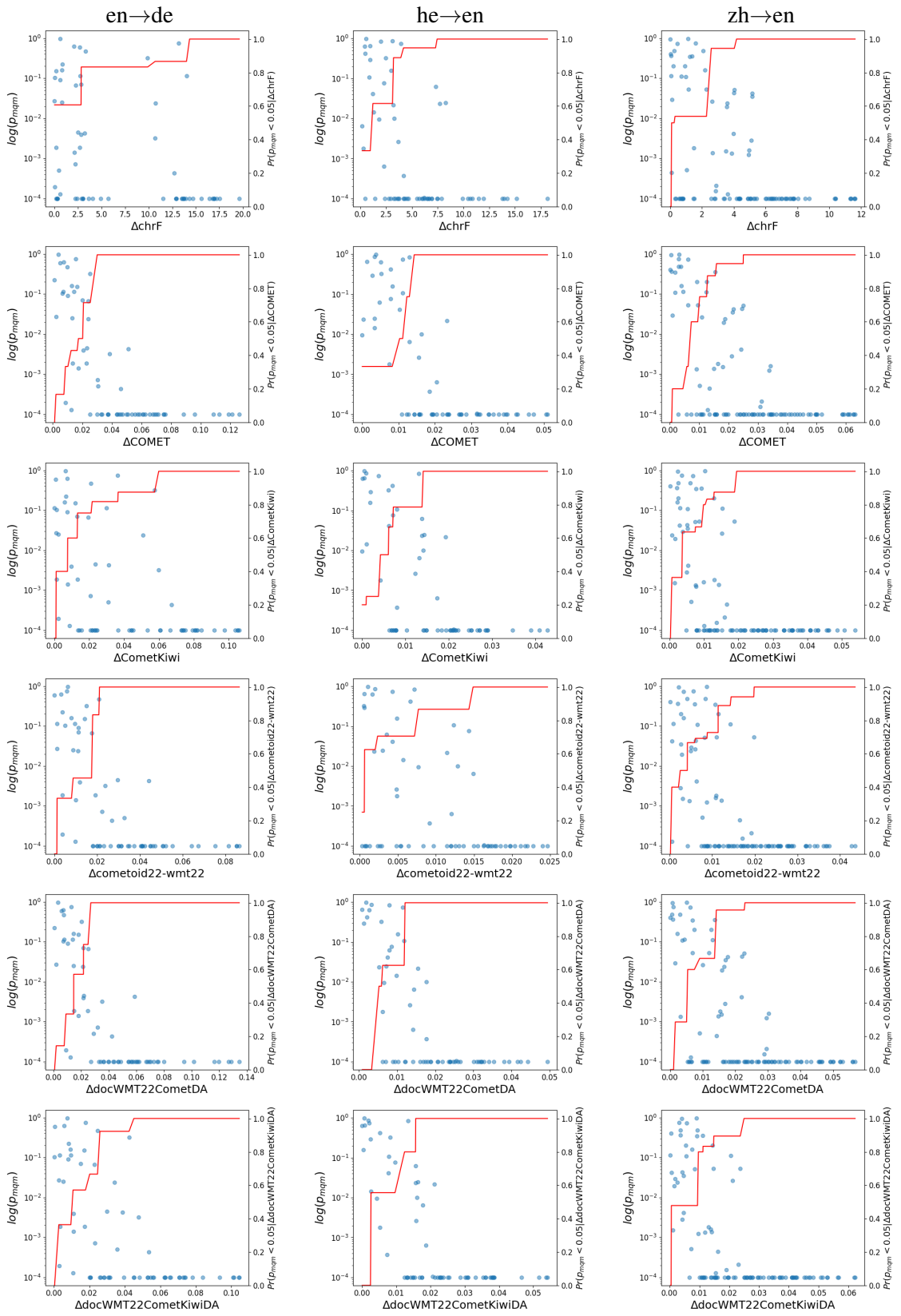


Figure 10: Log p-value of one-sided paired t-test on MQM scores (p_{mqm}) against the score difference of each metric (top to bottom: CHRF, COMET, COMETKIWI, COMETOID22-WMT22, DOCWMT22COMETDA, DOCWMT22COMETKIWIDA) for each system pair in each translation direction (left to right: en→de, he→en, zh→en). The red line is the isotonic regression fit to all data points, representing $Pr(p_{mqm} < 0.05|\Delta M)$. Note: for readability, values of p_{mqm} are rounded up to 0.0001 when they are less than 0.0001.

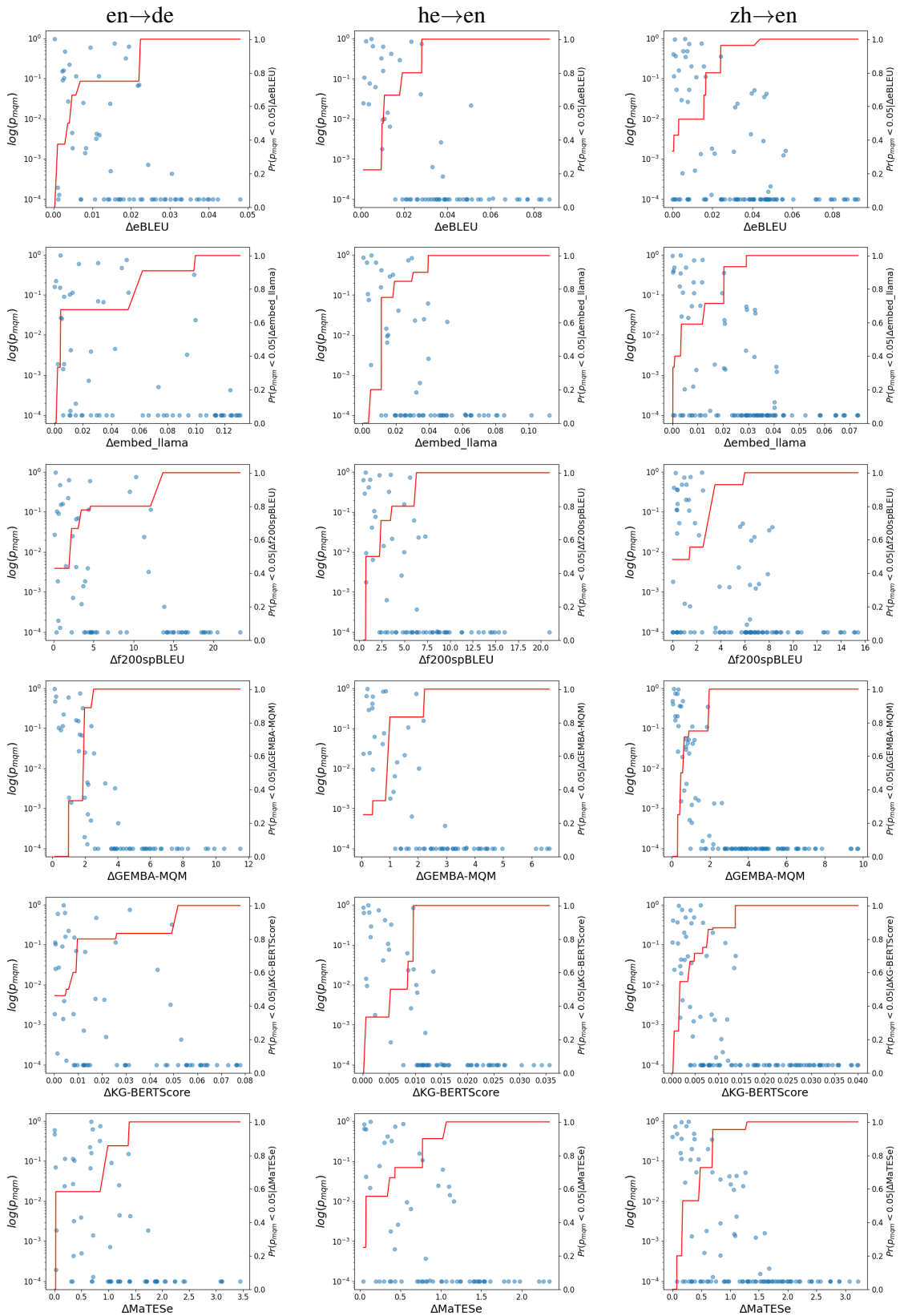


Figure 11: Log p-value of one-sided paired t-test on MQM scores (p_{mqm}) against the score difference of each metric (top to bottom: EBLEU, EMBED_LLAMA, F200SPBLEU, GEMBA-MQM, KG-BERTSCORE, MATESE) for each system pair in each translation direction (left to right: en→de, he→en, zh→en). The red line is the isotonic regression fit to all data points, representing $Pr(p_{mqm} < 0.05|\Delta M)$. Note: for readability, values of p_{mqm} are rounded up to 0.0001 when they are less than 0.0001.

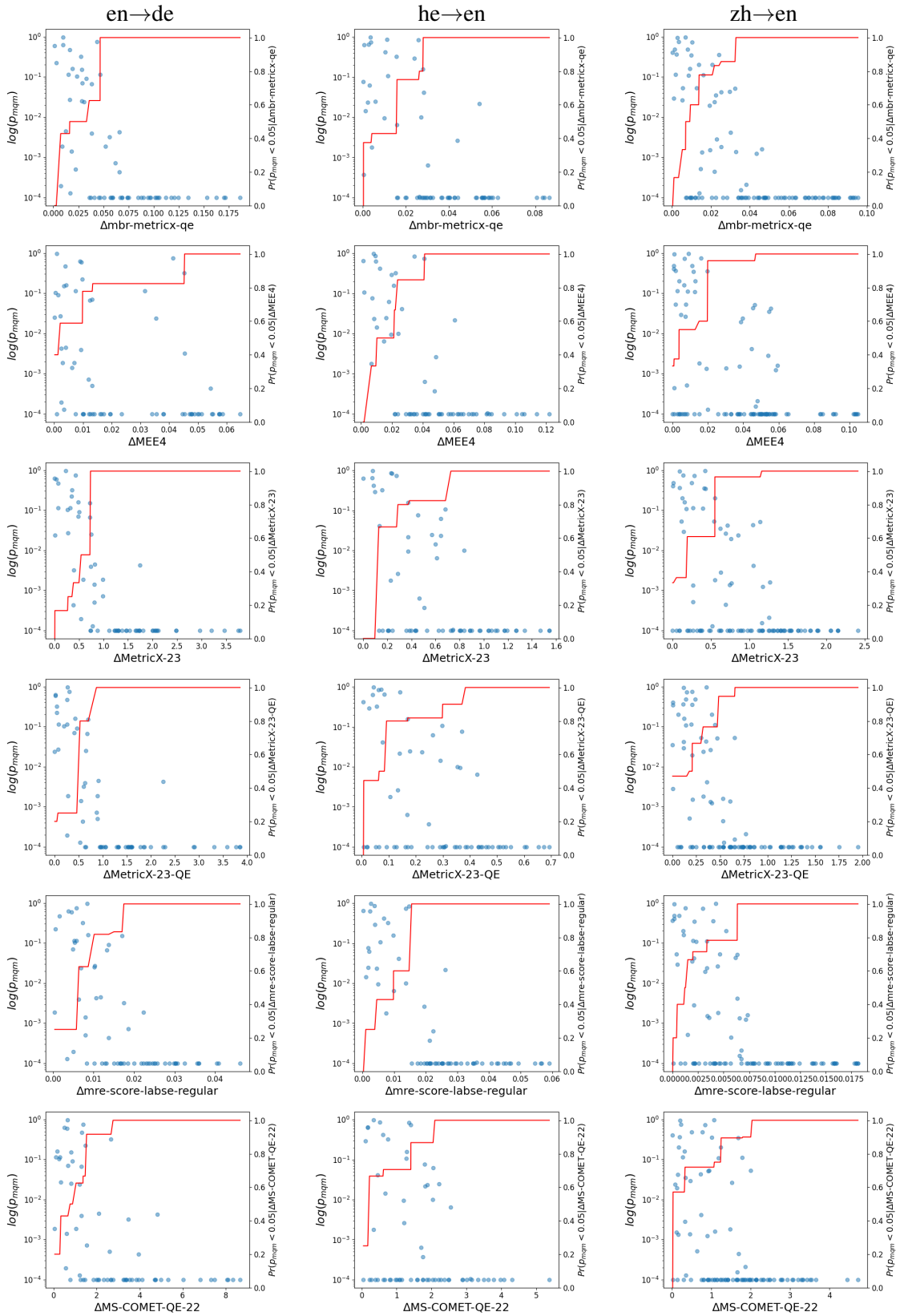


Figure 12: Log p-value of one-sided paired t-test on MQM scores (p_{mqm}) against the score difference of each metric (top to bottom: MBR-METRICX-QE, MEE4, METRICX-23, METRICX-23-QE, MRE-SCORE-LABSE-REGULAR, MS-COMET-QE-22) for each system pair in each translation direction (left to right: en→de, he→en, zh→en). The red line is the isotonic regression fit to all data points, representing $Pr(p_{mqm} < 0.05|\Delta M)$. Note: for readability, values of p_{mqm} are rounded up to 0.0001 when they are less than 0.0001.

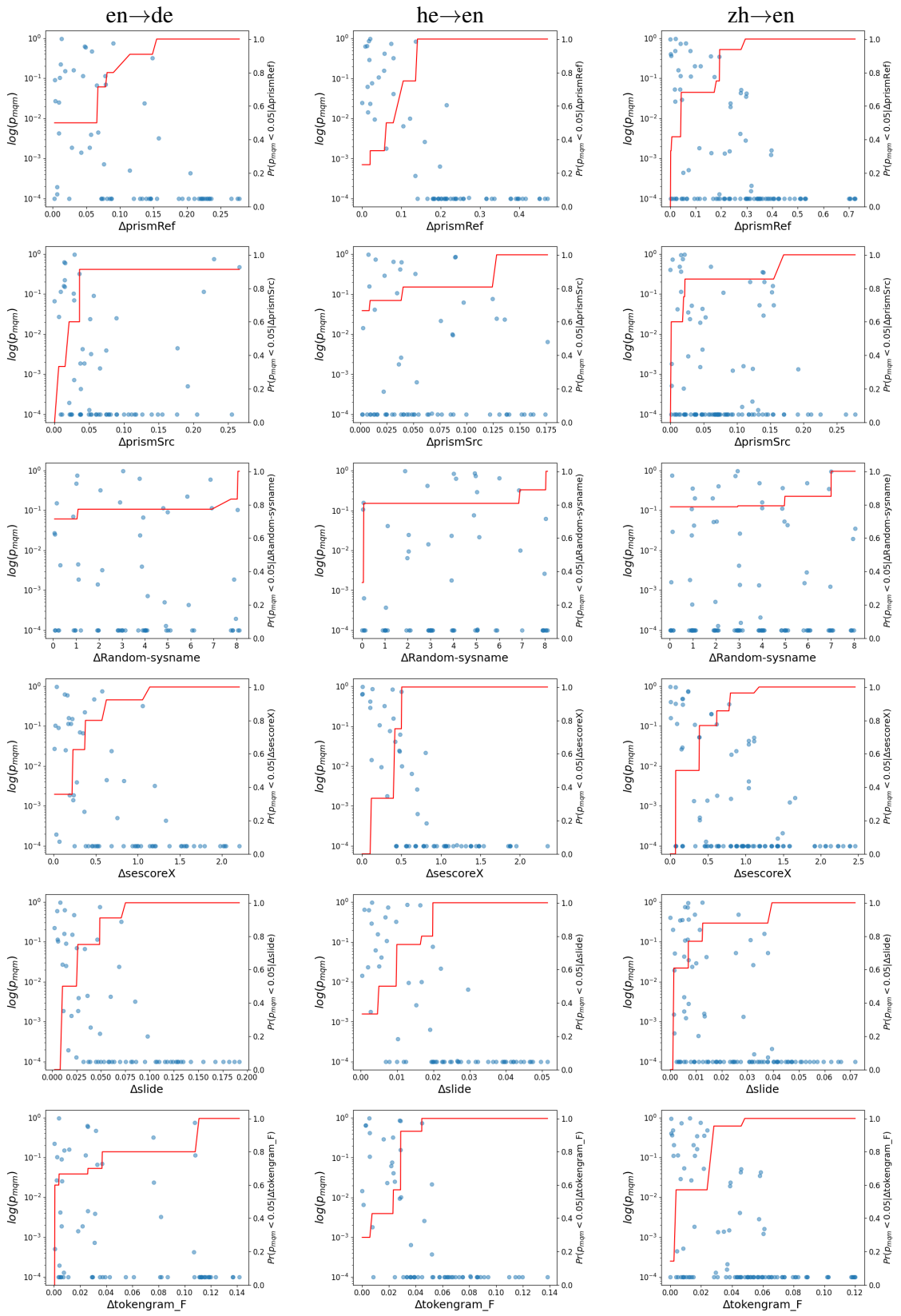


Figure 13: Log p-value of one-sided paired t-test on MQM scores (p_{mqm}) against the score difference of each metric (top to bottom: PRISMREF, PRISMSRC, RANDOM-SYSNAME, SESCOREX, SLIDE, TOKENGRAM_F) for each system pair in each translation direction (left to right: en→de, he→en, zh→en). The red line is the isotonic regression fit to all data points, representing $Pr(p_{mqm} < 0.05|\Delta M)$. Note: for readability, values of p_{mqm} are rounded up to 0.0001 when they are less than 0.0001.

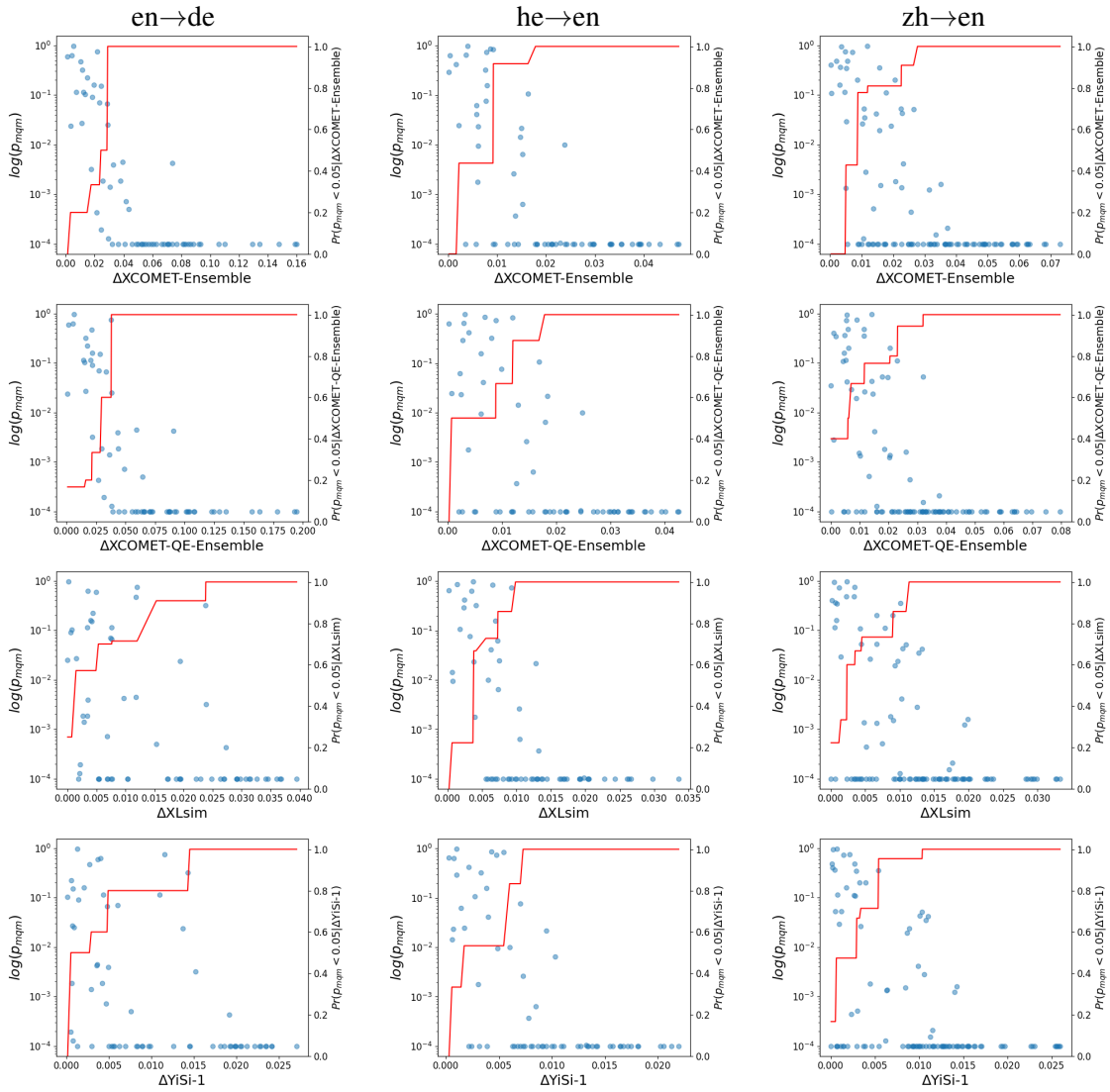


Figure 14: Log p-value of one-sided paired t-test on MQM scores (p_{mqm}) against the score difference of each metric (top to bottom: XCOMET-ENSEMBLE, XCOMET-QE-ENSEMBLE, XLSIM, YISI-1) for each system pair in each translation direction (left to right: en→de, he→en, zh→en). The red line is the isotonic regression fit to all data points, representing $Pr(p_{mqm} < 0.05|\Delta M)$. Note: for readability, values of p_{mqm} are rounded up to 0.0001 when they are less than 0.0001.

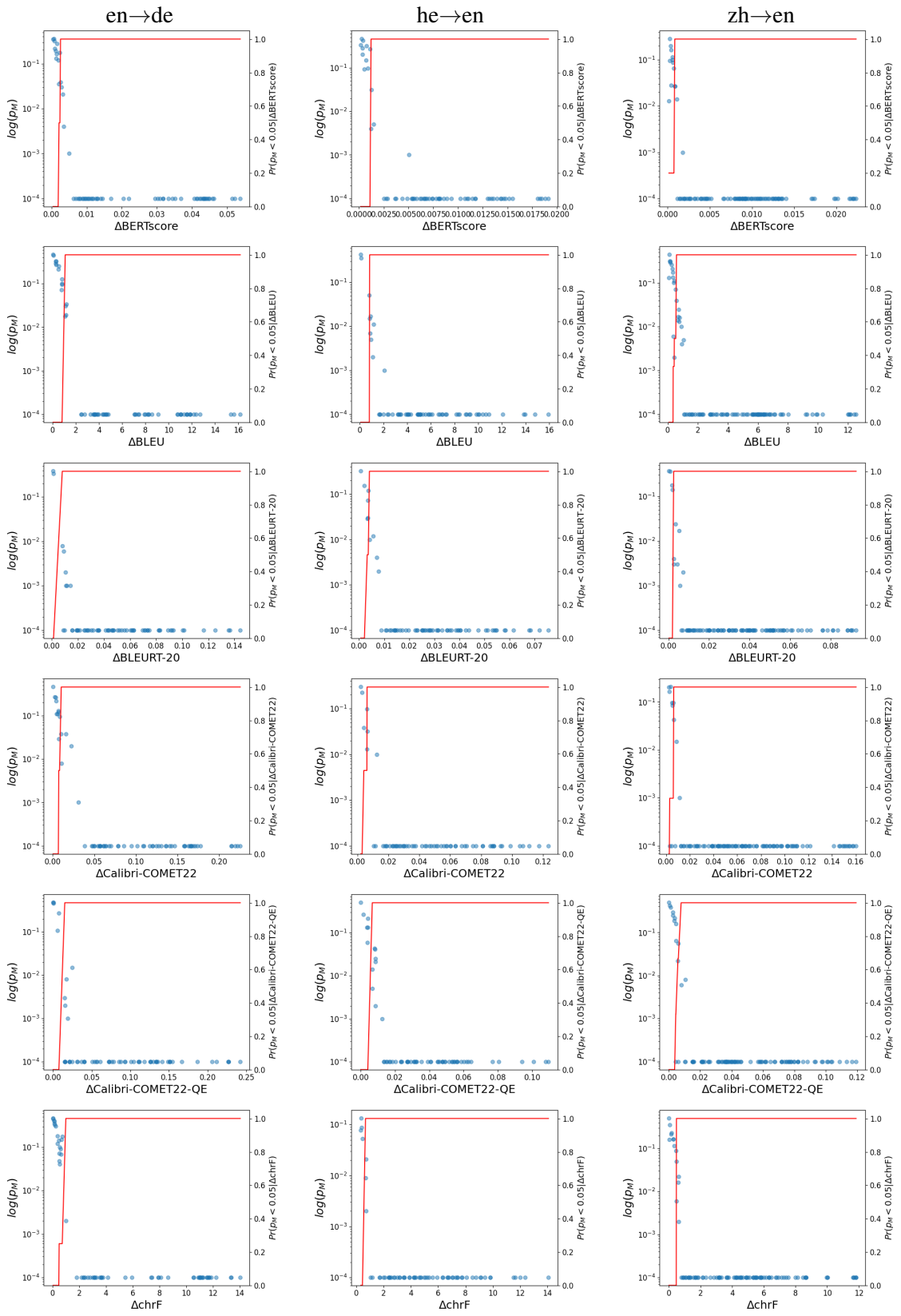


Figure 15: Log p-value of significance test with bootstrap resampling (p_M) on system-level metric scores against each metric (top to bottom: BERTSCORE, BLEU, BLEURT-20, CALIBRI-COMET22, CALIBRI-COMET22-QE, CHRF) score difference for each system pair in each translation direction (left to right: en→de, he→en, zh→en). The red line is the isotonic regression fit to all data points, representing $Pr(p_M < 0.05|\Delta M)$. Note: for readability, values of p_M are rounded up to 0.0001 when they are less than 0.0001.

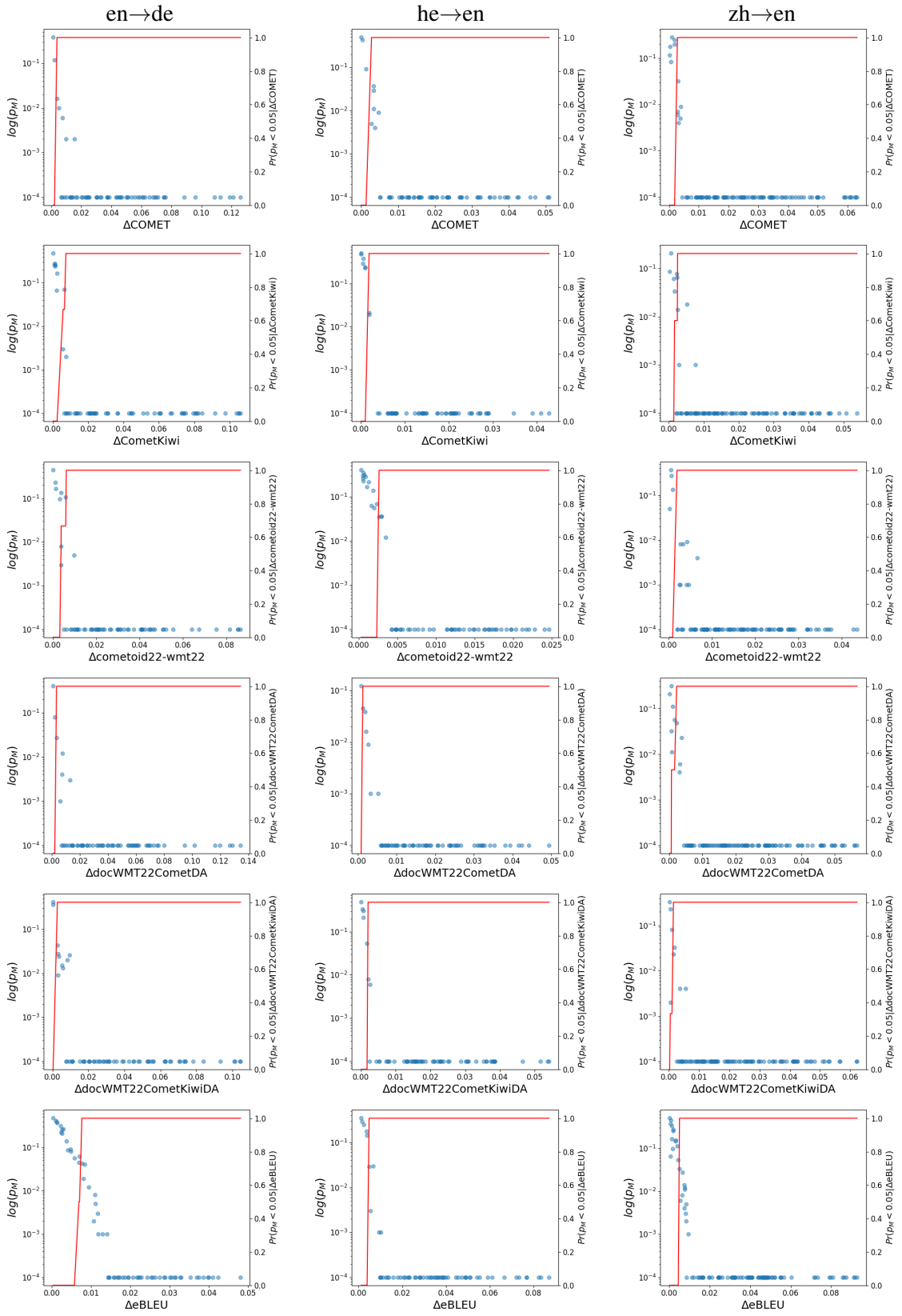


Figure 16: Log p-value of significance test with bootstrap resampling (p_M) on system-level metric scores against each metric (top to bottom: COMET, COMETKIWI, COMETOID22-WMT22, DOCWMT22COMETDA, DOCWMT22COMETKIWIDA, EBLEU) score difference for each system pair in each translation direction (left to right: en→de, he→en, zh→en). The red line is the isotonic regression fit to all data points, representing $Pr(p_M < 0.05|\Delta M)$. Note: for readability, values of p_M are rounded up to 0.0001 when they are less than 0.0001.

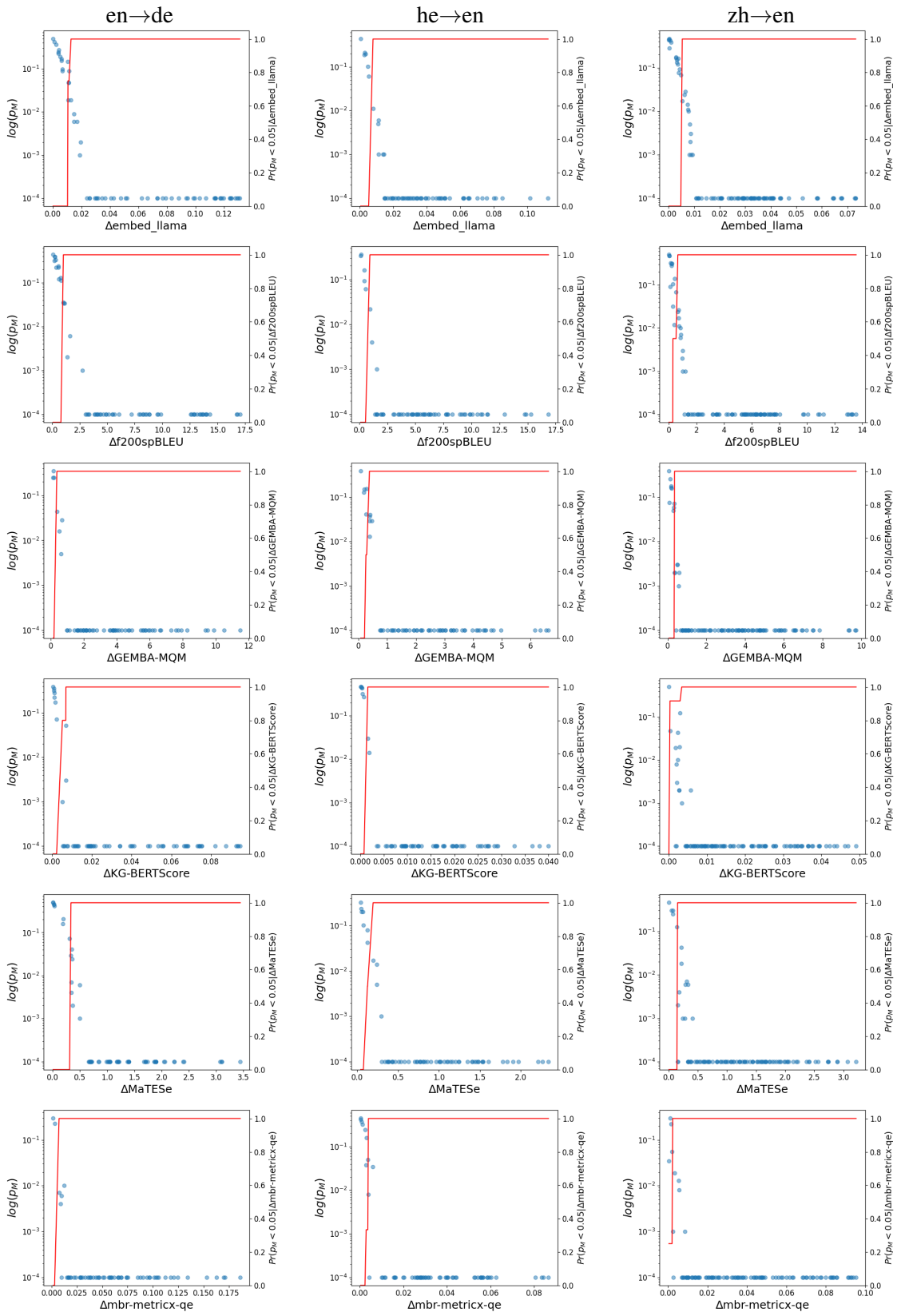


Figure 17: Log p-value of significance test with bootstrap resampling (p_M) on system-level metric scores against each metric (top to bottom: EMBED_LLAMA, F200SPBLEU, GEMBA-MQM, KG-BERTSCORE, MATESE, MBR-METRICX-QE) score difference for each system pair in each translation direction (left to right: en→de, he→en, zh→en). The red line is the isotonic regression fit to all data points, representing $Pr(p_M < 0.05|\Delta M)$. Note: for readability, values of p_M are rounded up to 0.0001 when they are less than 0.0001.

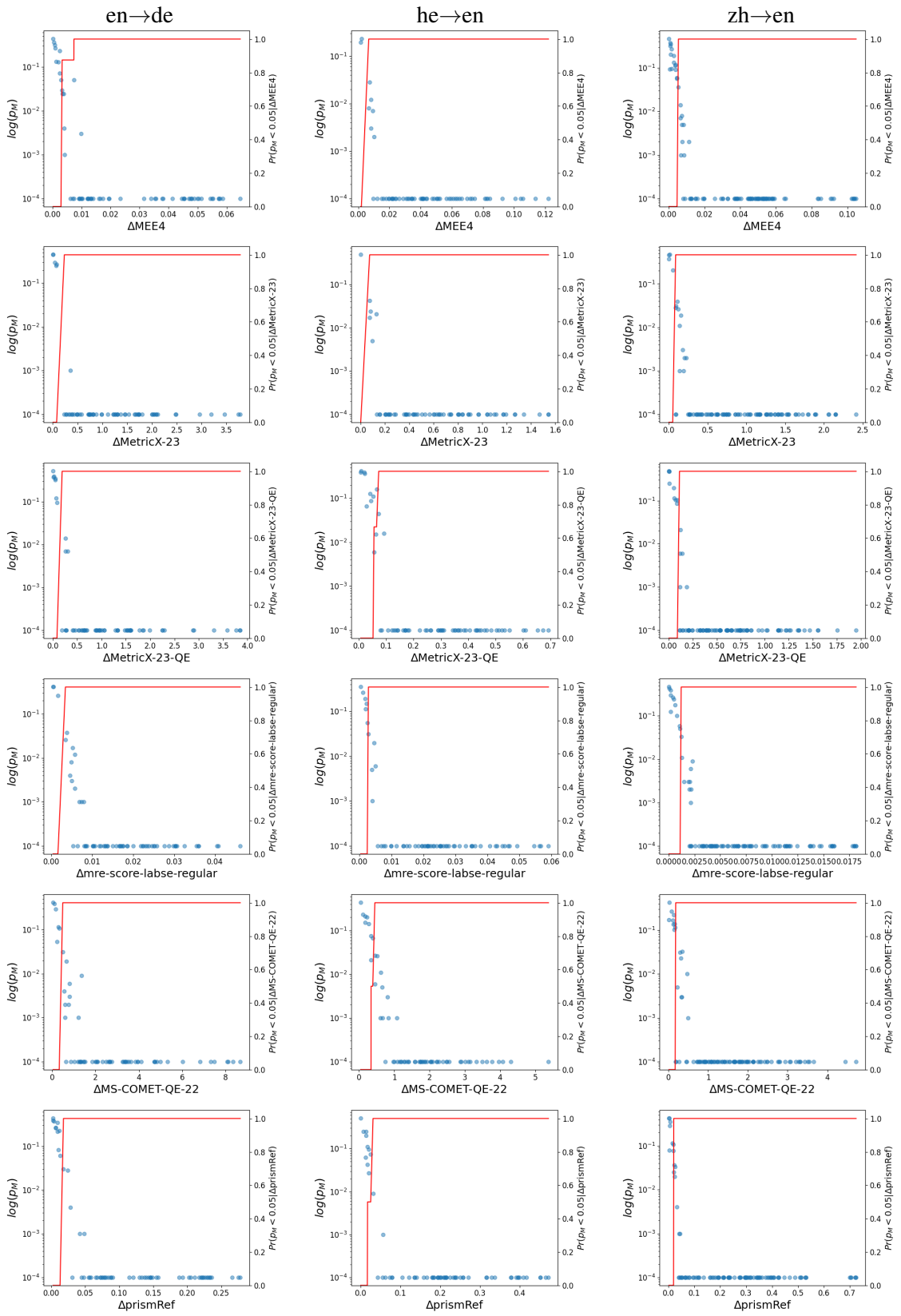


Figure 18: Log p-value of significance test with bootstrap resampling (p_M) on system-level metric scores against each metric (top to bottom: MEE4, METRICX-23, METRICX-23-QE, MRE-SCORE-LABSE-REGULAR, MS-COMET-QE-22, PRISMREF) score difference for each system pair in each translation direction (left to right: en→de, he→en, zh→en). The red line is the isotonic regression fit to all data points, representing $Pr(p_M < 0.05|\Delta M)$. Note: for readability, values of p_M are rounded up to 0.0001 when they are less than 0.0001.

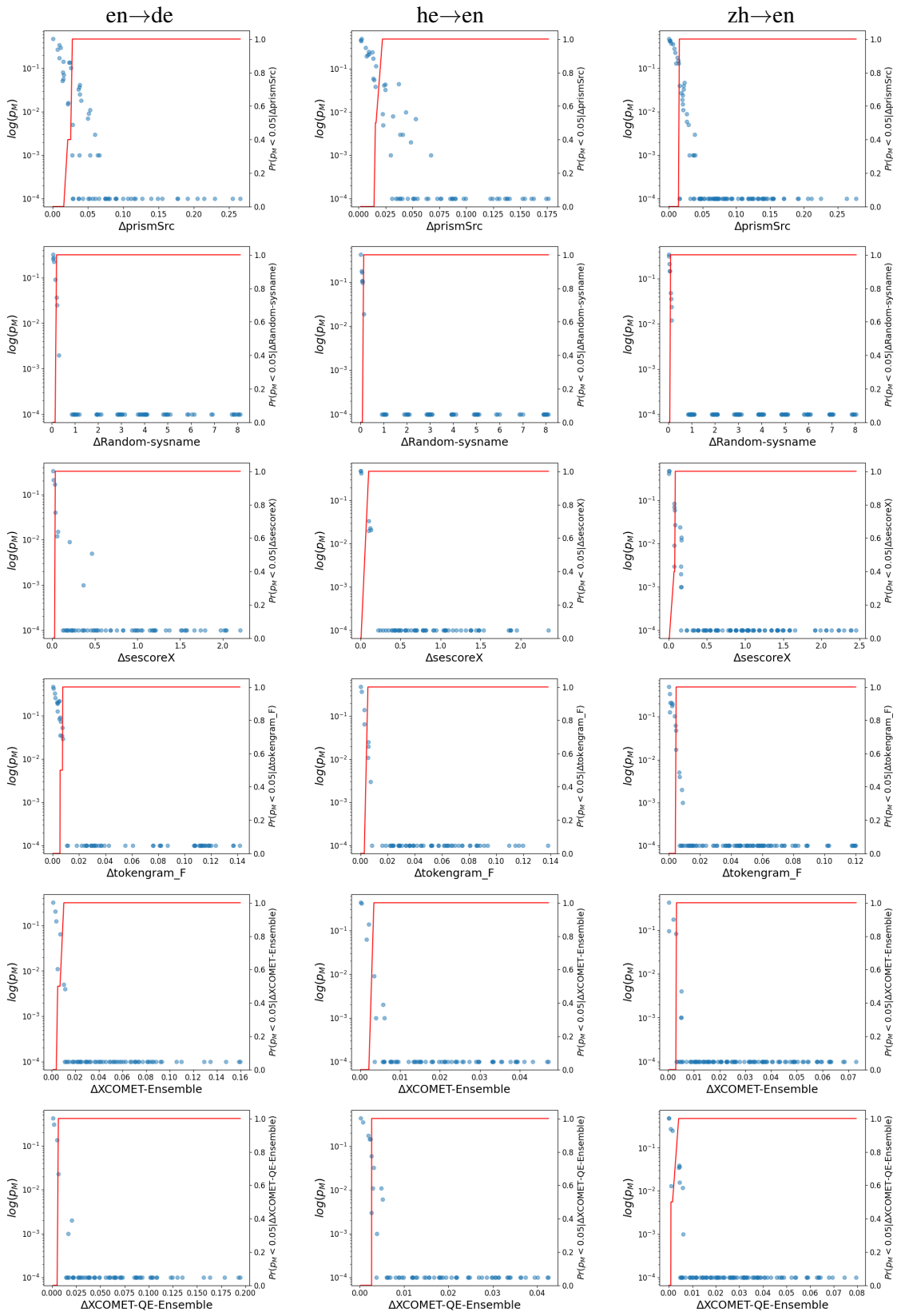


Figure 19: Log p-value of significance test with bootstrap resampling (p_M) on system-level metric scores against each metric (top to bottom: PRISMSRC, RANDOM-SYSNAME, SESCOREX, TOKENGRAM_F, XCOMET-ENSEMBLE, XCOMET-QE-ENSEMBLE) score difference for each system pair in each translation direction (left to right: en→de, he→en, zh→en). The red line is the isotonic regression fit to all data points, representing $Pr(p_M < 0.05|\Delta M)$. Note: for readability, values of p_M are rounded up to 0.0001 when they are less than 0.0001.

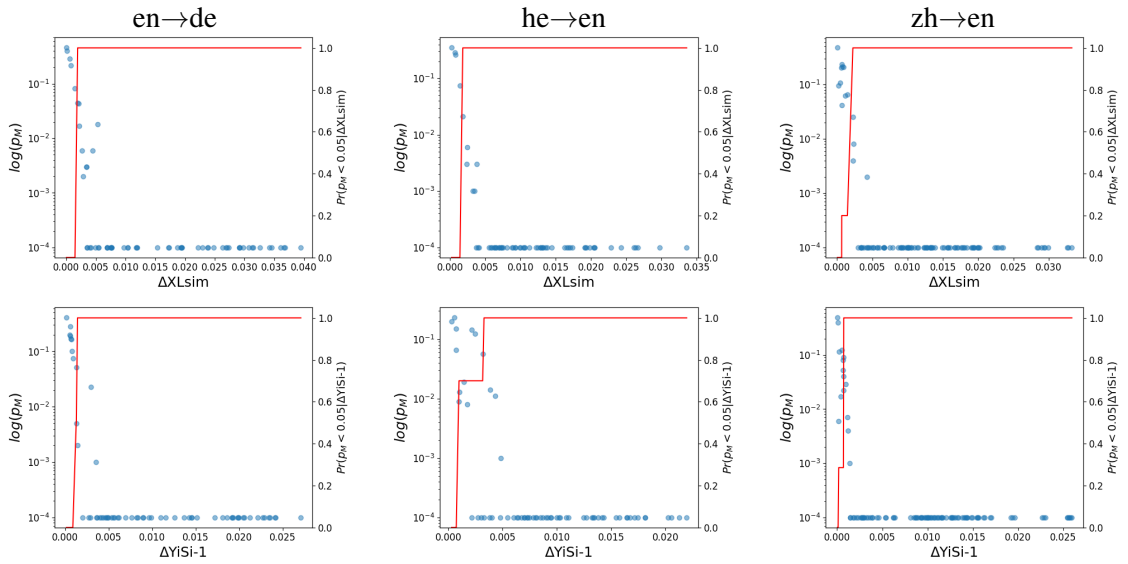


Figure 20: Log p-value of significance test with bootstrap resampling (p_M) on system-level metric scores against each metric (top to bottom: XLSIM, YISI-1) score difference for each system pair in each translation direction (left to right: en→de, he→en, zh→en). The red line is the isotonic regression fit to all data points, representing $Pr(p_M < 0.05|\Delta M)$. Note: for readability, values of p_M are rounded up to 0.0001 when they are less than 0.0001.

Findings of the WMT 2023 Shared Task on Quality Estimation

Frédéric Blain[(1)], Chrysoula Zerva[(2,3)], Ricardo Rei[(3,4,5)], Nuno M. Guerreiro[(2,3,4,8)], Diptesh Kanojia[(6)], José G. C. de Souza[(4)], Beatriz Silva[(4)], Tânia Vaz[(4)], Yan Jingxuan[(4)], Fatemeh Azadi[(7)], Constantin Orăsan[(6)], André F. T. Martins[(2,3,4)]

[(1)]Tilburg University, [(2)]Instituto de Telecomunicações, [(3)]Instituto Superior Técnico, [(4)]Unbabel, [(5)]INESC-ID, [(6)]University of Surrey, [(7)]University of Tehran, [(8)]MICS, CentraleSupélec, Université Paris-Saclay

f.l.g.blain@tilburguniversity.edu, {d.kanojia,c.orasan}@surrey.ac.uk, jose.souza@unbabel.com

ft.azadi@ut.ac.ir, {chrysoula.zerva,ricardo.rei,nuno.s.guerreiro,andre.t.martins}@tecnico.ulisboa.pt

Abstract

We report the results of the WMT 2023 shared task on Quality Estimation, in which the challenge is to predict the quality of the output of neural machine translation systems at the word and sentence levels, without access to reference translations. This edition introduces a few novel aspects and extensions that aim to enable more fine-grained, and explainable quality estimation approaches. We introduce an updated quality annotation scheme using Multidimensional Quality Metrics to obtain sentence- and word-level quality scores for three language pairs. We also extend the provided data to new language pairs: we specifically target low-resource languages and provide training, development and test data for English-Hindi, English-Tamil, English-Telegu and English-Gujarati as well as a zero-shot testset for English-Farsi. Further, we introduce a novel fine-grained error prediction task aspiring to motivate research towards more detailed quality predictions.

1 Introduction

This edition of the shared task on Quality Estimation (QE) aims to build on previous editions and findings to further benchmark methods for estimating the quality of neural machine translation (MT) output at run-time, without the use of reference translations. It includes (sub)tasks that consider the quality of machine translations at word- and sentence-level.

Over the past years, the QE field has been moving towards explainable, large, multilingual models that have been shown to achieve high performance, especially at sentence-level (Specia et al., 2021; Zerva et al., 2022). The recent proliferation of Large Language Model (LLM) technology and the consequential performance improvements in MT elevate the significance of advancing methodologies for quality estimation. In light of this, emphasis should be placed on multilingual quality estimation, in particular for low- and medium-resource languages, necessitating the development of more precise and interpretable quality assessment techniques. Additionally, it is important to address the challenge of robustness to hallucinations, prioritise sustainability, and optimise computational efficiency. These considerations collectively contribute to progress toward trustworthy and dependable QE systems that could facilitate real-time, reliable assessments of translation quality.

In this edition of the shared task, we further expand the provided resources, introducing new low-resource language pairs for Indian languages, namely Marathi, Tamil, Telugu, Gujarati and Hindi, as well as Farsi and Hebrew. Following the previous editions, we provide both annotations for direct assessments (DA), post-edits (PE) and Multidimensional Quality Metrics (MQM) (Lommel et al., 2014). We describe in detail the annotation process and provide statistics for the different language pairs in Section 2.

Overall, in addition to advancing the state-of-the-art at all prediction levels, our main goals are:

- to extend the languages covered in our datasets with low- and medium-resource languages;
- to investigate the potential of fine-grained quality estimation;
- to investigate new multilingual and language independent approaches esp. with regards to zero-shot approaches;
- to study the robustness of QE approaches to hallucinations; and
- to continue monitoring the computational efficiency of proposed approaches for sustainability purposes.

We thus designed two tasks this year:

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 629–653
December 6–7, 2023.

Task 1 The core QE task, which consists of separate sentence-level and word-level sub-tasks. For the sentence-level sub-tasks, the goal is to predict a quality score for each segment in a given test set, which can be a variant of DA (§2.1) or MQM (§2.2). For the word-level sub-task, participants had to predict translation errors in the form of binary quality tags (see §3.1.3).

Task 2 The fine-grained error prediction task, where participants were asked to detect error spans alongside error severities (§3.2).

The tasks make use of large datasets annotated by professional translators with either 0-100 DA scoring, post-editing or MQM annotations. We update the training and development datasets of previous editions and provide new test sets for Tasks 1 and 2. The datasets and models released are publicly available[1].

Apart from the data made available through the QE shared task, participants were also allowed to explore any additional data and resources deemed relevant, across tasks. In addition, LLMs could also be used both to extend resources and to complement predictions.

The shared task uses CodaLab as a submission platform, where each sub-task corresponds to a separate competition instance. Participants (Section 5) could submit up to a total of 10 submissions per sub-task. Results for all tasks evaluated according to standard metrics are given in Section 6. Baseline systems were trained by the task organisers and entered into the platform to provide a basis for comparison (Section 4). A discussion on the main goals and findings from this year's task is presented in Section 7.

2 Datasets

2.1 DA & Post-edit data:

For all language pairs, the data provided is selected from publicly available resources. Specifically for training, we used the language pairs from the MLQE-PE dataset (Fomicheva et al., 2022), as well as newly annotated data for languages spoken in India (Hindi, Tamil, Telugu and Gujarati). Overall, we provided training data for 15 language pairs with DA annotations, 12 with post edits, and 3 with MQM annotations, accounting for a mix of high, medium and low-resource languages. The statistics for the provided data are detailed in Table 1.

For the English-Marathi language pair included in the last edition, we provided a novel test set this year. To expand on language resources for the QE shared task, we chose Hindi (Hi) and Gujarati (Gu) as target languages from the Indo-Aryan language family, Tamil (Ta) and Telugu (Te) were chosen from the Dravidian language family. For En-Hi, En-Ta, En-Te, and En-Gu, dataset curation and annotation were performed with the help of professional translators who were native speakers of the target language. The annotators were provided with guidelines which discussed DA score ranges with various error types. Additionally, parallel segments were curated from the following parallel corpora: *i) Anuvaad* parallel corpus[2] (General, Healthcare and Legal domain; *ii*) IITB English-Hindi parallel corpus[3] (Kunchukuttan et al., 2018) (Culture/-Tourism domain), and parallel segments scraped from NPTEL[4]; and *iii*) SpokenTutorials[5] (Education domain). The curated segments were selected from the above-mentioned domains to ensure cross-domain impact and performance.

From the *Anuvaad* parallel corpus, we filtered source and parallel segments based on LaBSE (Feng et al., 2022) at high threshold values in the range [0.85, 0.99]. This helps us ensure the presence of good-quality reference translations from a noisy parallel corpus. We then selected source sentences for the dataset by varying token length in buckets of $0-10$, $10-20$, and $20-30$ tokens. This allows us to get annotations on various sentence lengths and helps manage the annotation cost to a certain extent. Moreover, translation models tend to generate erroneously over longer sequences (Varis and Bojar, 2021), and ensuring short and longer source sentences are a part of the data helps us presume a balanced DA distribution in the human annotation. We obtained the translation with the 1.3B parameter NLLB model (Costa-jussà et al., 2022) from HuggingFace[6]. The inference was performed with 5 beams, limiting the n-gram repetition to 2 and maximum length to 80 tokens, with early stopping enabled. The curation of source

[1] https://github.com/WMT-QE-Task/wmt-qe-2023-data

[2] https://github.com/project-anuvaad/anuvaad-parallel-corpus

[3] Unreleased parallel segments, to be released here in v3.2: https://www.cfilt.iitb.ac.in/iitb_parallel/

[4] https://nptel.ac.in/

[5] https://spoken-tutorial.org/

[6] https://huggingface.co/facebook/nllb-200-distilled-1.3B

segments from parallel corpora allowed us to compare the performance with IndicTrans (Ramesh et al., 2022) and 600M parameter NLLB model, in terms of TER and BLEU, helping us select the model and parameters above.

During the annotation, weekly validation of randomly selected instances was performed by an unbiased native speaker who provided feedback to further improve annotations during the data curation. After all three annotators performed the DA annotations, we separated the data into training, development, and test sets while filtering for a balanced distribution of DA scores across all sets.

For the En-Fa dataset, we used the post-edited data provided in Azadi et al. (2022) to get the word-level quality annotations. It contains 1K sentences derived from some English scientific articles in the domains of technology, computer science, and humanities. These sentences were firstly translated to Farsi, using an RNN-based commercial MT system named Faraazin[7]. Then, each sentence was given to a professional human translator to be post-edited and provide the correct translation with minimum edits. These post-edits were finally validated by another annotator to ensure their quality.

2.2 MQM Data

As **training data**, we used the annotations released for the Metrics and QE shared tasks in the previous years (Freitag et al., 2021a,b). Together, these annotations, cover 3 high-resource language pairs, namely: Chinese-English (Zh-En), English-German (En-De) and English-Russian (En-Ru), and span across two domains (News and Ted Talks).

As **test data**, we annotated new evaluation sets for three language directions. A low-resource language pair, Hebrew-English (He-En), and two high-resource language pairs, English-German and Chinese-English. The evaluation sets were annotated by professional translators following a MQM typology (Burchardt, 2013) and specific guidelines[8].

The documents used for the evaluation sets are shared with the General MT task in WMT and follow the same distribution of domains in that data. These documents were translated using the NLLB (Team et al., 2022) model of 1.3B parameters[9], the same model used in Section 2.1. We note that the En-De sources were originally organised in document-level, and we opted for converting them to smaller segments, so that we do not divert from the processing applied for the other LPs. Hence we first applied sentence splitting and then followed the same translation and annotation process described in this section.

All evaluation sets were annotated by professional translators and, for En-De and Zh-En the annotations were reviewed by a separate group of professional translators that amended any incoherences or disagreements from the first round of annotation. Regarding the domains of the data, for He-En, two different evaluation sets were annotated, one with newswire articles and another from product user reviews. For En-De, documents from four domains were annotated: transcriptions of meetings, newswire articles, social media posts, and product user reviews. For Zh-En, documents from three domains were used: manuals from information technology software or devices, newswire articles and product user reviews.

3 Quality Estimation tasks

In what follows, we briefly describe each sub-task, including the datasets provided for them.

3.1 Task 1: Predicting translation quality

The ability to accurately estimate the quality of translations on sentence- or word-level on-the-fly, i.e., without access to human-references is at the core of the QE shared task. Sentence and word-level estimates can provide complementary views of the quality of a sentence capturing different aspects (e.g. overall fluency versus specific mistranslations).

Following last edition, the data was produced in the following ways:

1. DA sentence level scores: The quality of each source-translation pair is annotated by at least 3 independent expert annotators, using DA on a scale 0-100.

2. MQM approach: Each source-translation pair is evaluated by at least 1 expert annotator, and errors identified in texts are highlighted and classified in terms of severity (minor, major, critical) and type (grammar correctness, omission, style, mistranslation, among others). We use this information for both word and sentence level quality scores.

[7] https://www.faraazin.ir/

[8] http://bit.ly/mqm-guidelines

[9] Model identifier FACEBOOK/NLLB-200-1.3B

3. Post-editing approach: The translation is post-edited to obtain the closest possible, fully correct translation of the source. By considering the alignment between the source, translation and post-edited sentence, we can propagate the errors from the source to the translated sentence and annotate the segments that were potentially mistranslated and/or not translated at all. We use this information to infer word-level quality scores.

The DA and MQM sentence level annotations were further processed to obtain normalised quality scores that have the same direction between high and low quality. We provide more details on the required pre-processing in §3.1.1 and §3.1.3.

3.1.1 Sentence-level quality prediction

This year we used a single competition instance both for DA and MQM-derived annotations aiming to motivate the submission of models that are robust to both annotation formats. To that end, we also aligned the scores by processing and normalising them as follows:

- For the **DA** scores we standardize the scores with respect to each annotator and then compute the mean average of standardized scores for each sentence.
- For the **MQM** scores we need to first compute the overall score from the individual errors. Hence for each annotator, we first compute the sentence-level score as

$$MQM^{sent}(hyp) = \frac{100 - \sum_{e \in hyp} severity(e)}{|hyp|}, \quad (1)$$

where hyp is a hypothesis sentence represented as a sequence of tokens, e is an error annotated in that sentence and the $severity$ is computed but adding:

  + 1 point for minor errors
  + 5 points for major errors
  + 10 points for critical errors

To align with DA annotations we subtract the summed penalties from 100 (perfect score) and we then divide by the sentence length (computed as number of words). We then normalise per annotator as in the DA case and compute the mean average in the case of multiple annotators.

Regarding evaluation, systems in this task (both for DA and MQM) are **evaluated against the true z-normalised sentence scores using Spearman's rank correlation coefficient ρ as the primary metric**. This is what was used for ranking system submissions. Pearson's correlation coefficient, r, and Kendall τ were also computed as secondary metrics but not used for the final ranking of systems.

3.1.2 Hallucinations

Hallucinations are highly pathological translations that contain content that is detached from the source (Raunak et al., 2021). As such, they can have devastating impact when models are deployed *in the wild* for real-world applications. Quality estimation systems are an appealing and attractive strategy to identify and flag these translations before they reach end-users. However, recent research has found that QE models may not appropriately penalize hallucinations and other critical errors (Raunak et al., 2022; Guerreiro et al., 2023c). This concern is further amplified for low-resource languages, where this undesirable behavior may arise even more frequently (Dale et al., 2023b). As such, in this edition of the shared task, we created data to assess the capability of submitted QE models in detecting hallucinations.

The data was created through a three-step process: (i) we started by generating translations for all language pairs of this year's shared task with NMT models[10], using the FLORES `devtest` and `test` splits (Goyal et al., 2022), as well as WikiMatrix data available through OPUS (Schwenk et al., 2019); then (ii) we automatically detected hallucinations generated by the models; and finally (iii) manually verified the flagged translations in order to guarantee that they are hallucinations. To automatically detect the hallucinations, we followed the procedure from Guerreiro et al. (2023a), which directly draws from several relevant contributions from research works in the literature of hallucination detection (Ferrando et al., 2022; Dale et al., 2023a; Guerreiro et al., 2023c).

To evaluate the performance of the submissions, we created, for each language pair, an *evaluation* set that consists of: all the hallucinations for the language pair, and the samples whose gold score is above the 25th percentile. This is to ensure that the non-hallucinations in the evaluation set are not

[10]We used the massively multilingual models (175M and 615M parameters) released in Goyal et al. (2022).

	Language Pairs	Sentences Train / Dev / Test23	Tokens Train / Dev / Test23	DA	PE	MQM	Data Source	Release
DA & post edits	En-De	10,000 / – / –	148,044 / – / –	✓	✓		Wikipedia	2021/22
	En-Zh	10,000 / – / –	148,529 / – / –	✓	✓		Wikipedia	2021/22
	Ru-En	10,000 / – / –	105,871 / – / –	✓	✓		Reddit	2021/22
	Ro-En	10,000 / – / –	154,825 / – / –	✓	✓		Wikipedia	2021/22
	Et-En	10,000 / – / –	126,547 / – / –	✓	✓		Wikipedia	2021/22
	Ne-En	10,000 / – / –	135,095 / – / –	✓	✓		Wikipedia	2021/22
	Si-En	10,000 / – / –	140,932 / – / –	✓	✓		Wikipedia	2021/22
	Ps-En	2,000 / – / –	54,459 / – / –	✓	✓		Wikipedia	2021/22
	Km-En	2,000 / – / –	44,029 / – / –	✓	✓		Wikipedia	2021/22
	En-Ja	2,000 / – / –	41,272 / – / –	✓	✓		Wikipedia	2021/22
	En-Cs	2,000 / – / –	40,638 / – / –	✓	✓		Wikipedia	2021/22
	En-Yo	1,010 / – / –	21,238 / – / –	✓	✓			2021/22
	En-Mr	27,000 / 1,000 / 1,086	717,581 / 26,253 / 27,951	✓	✓		multi-domain/multi-corpus	2022/23
	En-Hi	7,000 / 1,000 / 1,074	181,336 / 25,943 / 28,032	✓			multi-domain/multi-corpus	2023
	En-Gu	7,000 / 1,000 / 1,075	153,685/ 21,238 / 23,084	✓			multi-domain/multi-corpus	2023
	En-Ta	7,000 / 1,000 / 1,067	150,670 / 21,655/ 20,342	✓			multi-domain/multi-corpus	2023
	En-Te	7,000 / 1,028 / 1,000	147,492 / 20,686 / 22,640	✓			multi-domain/multi-corpus	2023
	En-Fa	– / – / 1,000	–/ – / 26,807		✓		news (multi-domain)	2023
MQM	En-De	30,425 / – / 1,897	877,066 / – / 37,996			✓	multi-domain	2021/23
	En-Ru	17,144 / – / –	395,045 / – / –			✓	multi-domain	2021/22
	Zh-En	36,851 / – / 1,675	1,654,454 / – / 39,770			✓	multi-domain	2021/23
	He-En	– / – / 1,182	– / – / 35,592			✓	multi-domain	2023

Table 1: Statistics of the data used for Task 1 and Task 2. The number of tokens is computed based on the source sentences. Hallucinated data included in the calculations for the 2023 testsets.

highly pathological translations (they may however be incorrect translations). We report the Area Under the Receiver Operating Characteristic curve (AUROC) and Recall at k (R@k), where k is defined as the number of hallucinations in the evaluation set. A perfect QE detector would have 100 AUROC and 100% Recall at k. We report the statistics of the evaluation sets in Table 8.

3.1.3 Word-level quality prediction

This sub-task focuses on detecting word-level errors in the MT output. The goal is to automatically predict the quality of each token using a binary decision, i.e., using OK as a label for tokens translated correctly and BAD otherwise.

We follow the annotation conventions of the previous edition, i.e., we do not consider source-side annotations, and incorporate omission errors to the target token annotations. Specifically, to account for omission errors, we consider the following convention: the token on the right side of the omitted text in the translation is annotated as "BAD". An additional `<EOS>` token is appended at the end of every translation segment to account for omissions at the end of each sentence. This allows the provision of a unified framework for both the post-edit originated annotations and the MQM annotations.

We thus use the same source-translation pairs used for the sentence-level tasks and obtain the binary tags as follows:

- For post-edited data, we use the methodology to obtain *translation error distance* (TER) scores (Snover et al., 2006) to obtain alignments between translation and post-edit and annotate the misaligned tokens as BAD.
- For MQM data, the tokens that fall within the text-spans annotated as errors (or any severity or category) are annotated as BAD. If the whitespace between two words is annotated as an error, then this is considered an omission, and the next token is annotated as BAD.

For the word-level task, **submissions are ranked using the Matthews Correlation Coefficient (MCC, Matthews, 1975) as the primary metric**, while F1-scores are provided as complementary information.

3.2 Task 2: Fine-grained error detection

For this task we attempt to focus on finer-grained quality predictions, taking advantage of the detailed information provided in the MQM annotation schema. Specifically, the MQM schema allows the annotation of additional information for each identified error. Specifically, each error span is annotated with error severity (*minor*, *major*, *critical*)

as well as error type (see also Figure 1). Such information allows for a more detailed analysis of the errors of MT systems, an understanding of their failure points and can provide the basis towards more explainable quality estimation.

Ideally, a fine-grained QE system is expected to be able to predict both the error type and its corresponding severity. However, the cardinality of error categories, the complexity of disentangling between them and the scarcity of MQM annotations render such a classification task particularly challenging. Hence, in this first attempt, we chose to focus only on **error severity** and merged together the major and the critical labelled errors due to the scarcity of the latter (see Table 2). As a result, we aimed to classify error spans as either *minor* or *major*.

LP	minor	major	critical
En-De	652	595	81
Zh-En	633	1063	242
He-En	792	1837	3

Table 2: Error severities (original: before merging critical and major severities) for the 2023 MQM test set.

As such, the information used for this task consists of: *i*) start and end index positions for each error span; and *ii*) the simplified error severity. The error spans are identified as sequences of continuous characters within a target hypothesis, allowing for annotations of single white spaces and punctuation marks in order to account for omission and punctuation errors respectively. Aiming to mimic the human annotations and simplify the task, overlapping error spans were allowed. Figure 1 shows an example of annotations.

For the evaluation, the primary metric was **F1-score**, computed on the character level and weighted to allow for half points for correctly identified span but misclassified severity. Precision and recall were also provided as complementary metrics. The evaluation approach is inspired by (Fonseca et al., 2019) but does not consider document-level annotations. With respect to overlapping annotations, we allow for multiple character level annotations [11] and will consider the best matching annotation per character position. As such for each segment we compute recall for the characters in

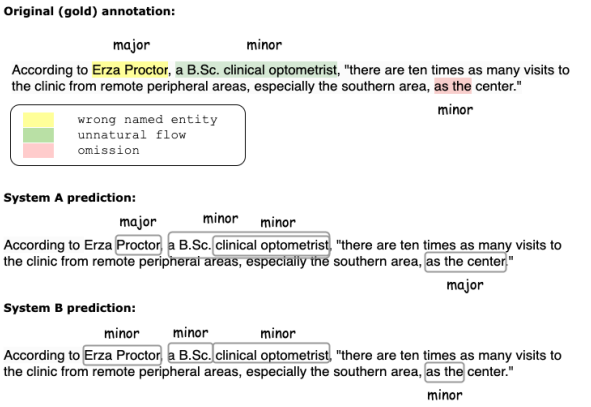


Figure 1: Example of gold annotations (MQM) for Task 2 (top) and respective prediction examples (bottom). Example taken from He-En test set.

gold annotation text spans, by computing the ratio between the overlap with system error spans and the gold error span length and weighting severity mismatches by 0.5. Respectively, we compute precision with respect to the system error span length and apply the same weighting convention (down-weighting by 0.5 for mismatched error severities). Figure 1 and Table 3 shows an example of the aforementioned process [12].

4 Baseline systems

4.1 Task 1: quality estimation

For the **sentence-level** sub-task, we opted for COMET-QE models (Rei et al., 2021) respectively pretrained on the DA and MQM QE data from WMT'21. Models are publicly available to download[13].

For the **word-level** sub-task, we trained a simplified architecture inspired by COMETKIWI (Rei et al., 2022a). More specifically, we used the multi-task architecture combining the sentence-level target and the binary word-level targets. However, we did not pretrain on HTER scores or Metrics data, and we skipped the few-shot language adaptation and language-specific tuning of task weights. The architecture of the baseline model is shown in Figure 2. The list of hyperparameters and their corresponding values can be found in appendix A.

[11] The gold data was processed to remove identical segments that correspond to the same span but have different error categories, but it preserved any partially overlapping segments that correspond to different error categories and/or severities.

[12] The link to evaluation scripts can be found at: https://github.com/WMT-QE-Task/qe-eval-scripts/blob/main/wmt23/

[13] https://wmt-qe-task.github.io/subtasks/task1/

Systems	Precision	Recall	F1-score
System A	$\frac{1*7+1*28+0.5*6}{7+28+13} = 0.79$	$\frac{1*7+1*28+0.5*6}{12+28+6} = 0.83$	0.81
System B	$\frac{0.5*12+1*28+0.5*6}{12+28+6} = 0.80$	$\frac{1*12+1*28+0.5*6}{12+28+6} = 0.80$	0.80

Table 3: Example of Precision and Recall computations for each annotation in the example of Figure 1.

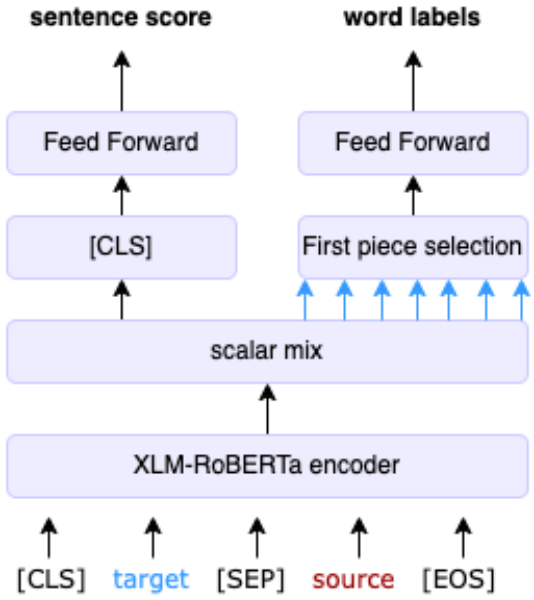


Figure 2: Baseline model for Task1 word level. Figure adapted from (Rei et al., 2022a)

4.2 Task 2: fine-grained error detection

For **Task 2** we constructed a simple baseline by using the same model used for Task 1 word-level, and post-process the predictions as follows:

- Detokenize the sentence
- Annotate continuous BAD tokens as a single text span
- Assume all errors are major

For all models, a large XLM-RoBERTa pre-trained encoder was used, without additional language tuning. The specific hyperparameters used are presented in Table 11.

5 Participants

Table 4 lists teams who officially took part to the QE shared task this year. In the remaining of this section, we report a brief system description gathered from each participant. For each team, we indicate the task(s) and sub-task(s) (*i.e.* language-pair(s)) they participated into.

Bering Lab (T1-SL; En-De, Zh-En): For each language pair, the team used an ensemble of the best three models from a pool of 10 models jointly trained for the word and sentence level tasks using a novel relative ranking loss function and Adversarial Weight Perturbation (AWP) to improve the robustness of the model. Using no additional pseudo-generated data, they pre-train the models using publicly available data from the previous WMT conferences that were augmented using the TER tool to generate binary word tags. They then fine-tuned 10 separate models on the labelled data from the WMT 2022 QE task randomly split into 10 folds. The models are fine-tuned in two steps. In the first step, the models are fine-tuned without AWP using the same objective as the pre-training step. Then, the best checkpoint from the first step is selected and tuned on the same objective, but with AWP. For the final submission they ensemble (average) the z-normalized scores from the top three models to get the final predictions.

(T1-WL; En-De, Zh-En): For each language pair, they use an ensemble of five models jointly trained for the word and sentence level tasks using a novel relative ranking loss function. Using no additional pseudo-generated data, they pre-train the models using publicly available data from the previous WMT conferences that were augmented using the TER tool to include word tags. They then split the labelled data from the WMT 2022 QE task into 20 folds and chose the best combination of five folds based on the minimum mean of the Kolmogorov-Smirnov goodness-of-fit scores between each validation set. Using these 5 folds, they fine-tune 5 final models using the same objective as the pre-training step. For the final prediction, the team chose the max score from each model for each token to get the final predictions.

NJUNLP (T1 & T2; En-De): Inspired by Direc-

tQE (Cui et al., 2021) and CLQE (Geng et al., 2023b), NJUNLP submission continues exploring pseudo data methods for QE. They generate pseudo MQM data using parallel data from the WMT translation task. Specifically, they replace the reference tokens with these tokens sampled from translation models. To simulate translation errors with different severities, they sample tokens with lower generation probabilities for worse errors. They pre-train the XLMR large model on pseudo MQM data, then fine-tune it on real QE data (including PE data). At both stages, they jointly learn sentence-level scores (MSE loss and margin ranking loss) and word-level tags (cross-entropy loss). For Task 1, the QE model outputs the sentence scores and the 'OK' probability of each token.

For Task 2, they set different thresholds for the 'OK' probability to predict fine-grained severity. They regard consecutive 'BAD' tokens as a whole span and take the worse severity of each token as the result. They train different models with different parallel data and ensemble their results as the final submission.

HW-TSC (T1-SL; En-De, Zh-En, En-Mr, En-Hi, En-Ta, En-Te, En-Gu): HW-TSC uses CrossQE, the same model as the one reported in (Tao et al., 2022), which consists of a multilingual base model and a task-specific downstream layer. The input is the concatenation of the source and the translated sentences. To enhance the performance, they finetuned and ensembled multiple base models using multilingual encoders such as XLM-RoBERTa, InfoXLM and RemBERT as well as a CometKiwi model. Moreover, they introduce a new corruption-based data augmentation method, which generates deletion, substitution and insertion errors in the original translation and uses a reference-based QE model to obtain pseudo scores.

(T2; all): For Task 2 they they convert the original word-level binary classification to a 3-way classification to adapt to Task 2 severities (no-error, minor, major). They then use a multitasking Comet model based on CometKiwi (Rei et al., 2023) which combines sentence scores and word-level tags using a weighted loss function. They set the weight of the sentence score sub-module to 0. They use InfoXLM-large and XLM-RoBERTa-large as the pre-trained encoders used during training and train on different data subsets for each LP. They finally use CometKiwi-da and continue to train a model based on CometKiwi-da. They finally combine the results over five checkpoints using the union of the predicted spans, which out-performed token-level majority voting.

KUNMT (T2; En-De, Zh-En): KUNMT proposes the use of different models to decompose tasks and post-editing with a large language model. In the process of error determination, span extraction, and severity assessment for each error span, distinct models were employed sequentially. The error determination model determines if an error exists in the sentence, and then the span assessment model explores the parts of the sentence where the error exists. For the spans where the error exists, the severity evaluation model evaluates whether the severity of the error is minor or major. All models were built upon XLM-RoBERTa-large, with some incorporating prompt-based learning. Results were subsequently calibrated using a large language model and tailored prompt engineering for the specific task.

Unbabel-IST (T1 & T2; all): the submission for Task 1 (word-level and sentence-level) follows their work from last year (Rei et al., 2022b). The major difference is the inclusion of the data from this year (e.g. sentence-level DA's for En-Te, En-Hi, En-Gu, En-Ta) and scaling the size of the pretrained encoder from InfoXLM to XLM-R XL and XXL (XXL was only used for sentence-level). They ensemble multiple checkpoints for the sentence-level subtask, using a weighted averaging of the predicted scores, optimised by LP.

For Task 2 they experimented with word-level models from Task 1 with GPT-4 prompts and with XCOMET (Guerreiro et al., 2023b). Their primary submission uses XCOMET which stands for eXplainable COMET. This model is trained with references to perform regression and error span identification. During inference the model can be used without refer-

ences, yet, for this task they found that using pseudo-references yields better performance if used with a simple heuristic where they first use a sentence-level QE system trained for Task 1 to evaluate the pseudo-reference. If the pseudo-reference is of high-quality, they give more weight to it otherwise, they give more weight to the source.

IOL Research (T1-SL; all): The IOL team experimented with several pretrained language models with extra modules to predict sentence level score and word tags including mBERT, XLM-RoBERTa-large, mDeberta, RemBert and InfoXLM. They first finetuned these models on DA and MQM scores data of QE and Metrics tasks in the previous years. Then, source text and its translation are fed into finetuned models added with extra modules for both sentence and word-level tasks. For sentence level, they separate embeddings of source text and translation of each layer in transformer models, and make a weighted sum among different layers for source and translation. Then the weighted embeddings of source and translation are concatenated and fed into a two-layer deep neural network to get score prediction with mean squared error (MSE) loss.

(T1-WL; all): For the word-level subtask, they use BiLSTM layer or one-layer DNN to do tag prediction on each token of translation with cross-entropy loss. The best checkpoint of each model is chosen by determining which checkpoint is best with respect to either Spearman correlation coefficient or MCC score, after training for three (3) epochs. The model of each language pair is tuned individually. The final result for each language pair is predicted by a weighted ensemble of different model checkpoints with LP-specific weights computed through weight searching using Optuna.

MMT (T1-SL; En-De, En-Mr, En-Hi, En-Ta, En-Te, En-Gu): For the studied language pairs, the MMT team enriched the training dataset through the application of eleven distinct data augmentation techniques, such as synonym substitution and back-translation, individually on the source sentence of each training instance. The results were generated using the best-performing model chosen from those trained on the corresponding augmented training datasets (in the case of English-German, the chosen model was trained on the augmented dataset created by applying the top four effective data augmentation techniques to each source sentence). The training methodology adheres to the COMET framework, with the foundational pre-trained model being XLM-RoBERTa-large.

SurreyAI (T1-SL; En-Mr, En-Hi, En-Ta, En-Te, En-Gu): The team proposes ensembleTQ as the main model, for which they train multiple multilingual QE models by fine-tuning pretrained language models (PTLMs) with autoencoder architecture. To that end they use the MonoTransQuest (Ranasinghe et al., 2020) architecture and report mean z-scores. The PTLMs that they combine in the ensemble are InfoXLM-large, XLMV-base and XLM-RoBERTa-large.

6 Results

In this section, we present and discuss the results of our shared task. Please note that for all the three sub-tasks we used statistical significance testing with $p = 0.05$.

6.1 Task 1

As we have seen in Task 1 description sentence-level submissions are evaluated against the true z-normalised sentence scores using Spearman's rank correlation coefficient ρ along with the following secondary metrics: Pearson's correlation coefficient, r, and Kendall's τ. Nonetheless, the final ranking between systems is calculated using the primary metric only (Spearman's ρ). Statistical significance was computed using William's test.

For the word-level task, the submissions are ranked using the Matthews correlation coefficient (MCC). F1-scores are provided as complementary information only and statistical significance was computed using randomisation tests (Yeh, 2000) with Bonferroni correction (Abdi, 2007) for each language pair.

The results for Task 1 are described in Tables 5 and 6. Looking at the obtained scores, we can observe an overall improved performance for the sentence-level scores, compared to previous years. While it is hard to make direct comparisons since

ID	Affiliations	
BeringLab	BeringLab	–
HW-TSC	Huawei Translation Services Center, China	(Li et al., 2023)
IOL Research	Transn IOL Research, China	(Yan, 2023)
KUNMT	Korea University, South Korea	–
MMT	University of Manchester, UK & ASUS Intelligent Could Services, Singapore & University of Melbourne, Australia	(Wu et al., 2023)
NJUNLP	Nanjing University & Huawei Translation Services Center, China	(Geng et al., 2023a)
Surrey AI	University of Surrey & Aston University, UK	(Sindhujan et al., 2023)
Unbabel-IST	Unbabel & INESC-ID & Instituto de Telecomunicações & Instituto Superior Técnico, Portugal	(Rei et al., 2023)

Table 4: Participants to the WMT23 Quality Estimation shared task.

the test-sets are new (and many language pairs are introduced for the first time), we can see that top performers obtain higher scores for En-Mr and Zh-En compared to the previous edition, and only for En-De we observe a relative drop (potentially justifiable by the introduction of more domains in the test set this year). Interestingly, for the word-level scores we observe higher correlations for the zero-shot tasks, as opposed to the ones where more training and development resources were made available.

We observe that especially for the sentence-level task all participants this year submitted ensembled predictions, with the ensemble size ranging from 3 to 12 models. We note that several teams combined models using different pre-trained encoders (HW-TSC, IOL Research, Surrey) and some participants focused their efforts on optimising the ensembled scores. Most notably, Bering Lab use a multi-step training where they select the best models on the first step, retrain with their proposed Adversarial Weight Perturbation method and then ensemble the top-3 models for the final submission. IOL research and Unbabel-IST also optimise the ensemble weights using optuna search.

We finally observe that following the trend of previous editions several participants experiment with training data augmentation techniques. Notably, most approaches this year focus on data augmentation that relates to the word-level or fine-grained annotations, either by computing TER-based word tags (Bering Lab), or by corrupting the target translations to generate pseudo-data with artificially generated error spans (NJUNLP, HW-TSC). For the latter, NJUNLP replace tokens and make use of the token distribution to approximate major versus minor errors (i.e, lower versus higher generation probabilities) and generate MQM-style annotations. Instead, HW-TSC propose to randomly corrupt the target (where corruption corresponds to insertion, deletion or replacement of a token) and use a heuristic score of the corrupted target to approximate a DA annotation style.

Best performers A total of seven teams participated in the **sentence-level sub-task**, yet only Unbabel-IST and IOL Research participated for all language pairs (including the zero-shot language pair, *He-En*), with Unbabel winning in the multilingual setting. However, for the individual language pairs, we observe different teams ranking at the top for different language pairs. Specifically, HW-TSC ranks at the top for all Indic language pairs, sharing the win with Unbabel-IST for En-Mr, En-Hi and En-Gu. On the MQM annotations, Unbabel-IST won the Zh-En and He-En language pairs, while IOL-Research and NJUNLP ranked top for En-De.

A total of four teams participated in the **word-level sub-task**, and similarly to the sentence-level only Unbabel-IST and IOL Research participated for all language pairs (including both zero-shot language pairs: *He-En* and *En-Fa*). NJUNLP won the task for the En-De language pair while Unbabel-IST ranked at the top for Zh-En, He-En, En-Mr and the multilingual task. IOL Research tied at the top for En-Fa.

We observe that while submissions consist of a mix of monolingual and multilingual submissions, and participants adopted a set of different strategies to design their architectures and tuning process, the top-ranking participants do share some common methodological choices. Specifically, all aforementioned participants tune their models in a multitasking setup, taking advantage of not only the sentence-level scores but also the word-level

Model	Multi	Multidimensional Quality Metric (MQM)			Direct Assessment (DA)				
		En-De	En-Zh	He-En	En-Mr	En-Hi	En-Ta	En-Te	En-Gu
Unbabel-IST	**0.594**	0.456	**0.493**	**0.668**	**0.704**	0.598	0.739	**0.388**	**0.714**
IOL Research	0.556	**0.483**	**0.482**	0.575	0.505	0.600	0.740	0.376	0.695
BASELINE	0.372	0.340	0.447	0.475	0.392	0.281	0.507	0.193	0.337
HW-TSC	–	0.437	0.460	–	**0.692**	**0.644**	**0.775**	**0.394**	0.691
MMT	–	0.316	–	–	0.650	0.494	0.547	0.337	0.540
SurreyAI	–	–	–	–	0.596	0.551	0.674	0.349	0.649
BeringLab	–	0.380	0.384	–	–	–	–	–	–
NJUNLP	–	**0.479**	–	–	–	–	–	–	–

Table 5: Spearman correlation for the official submissions to WMT23 Quality Estimation **Task 1 Sentence-level**. For each language pair, results marked in bold correspond to the winning submissions, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey

Model	Multi	Multidimensional Quality Metric (MQM)			Post-Editing (PE)	
		En-De	Zh-En	He-En	En-Mr	En-Fa
Unbabel-IST	**0.329**	0.246	**0.302**	**0.402**	**0.347**	0.345
IOL Research	0.298	0.256	0.250	0.359	0.334	**0.351**
BASELINE	0.252	0.179	0.225	0.275	0.287	0.293
BeringLab	–	0.233	0.241	–	–	–
NJUNLP	–	**0.297**	–	–	–	–

Table 6: Matthews Correlation Coefficient (MCC) for the official submissions to WMT23 Quality Estimation **Task 1 Word-level**. For each language pair, results marked in bold correspond to the winning submissions, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey

Model	Multi	Multidimensional Quality Metric (MQM)		
		En-De	En-Zh	He-En
Unbabel-IST	**0.220**	0.273	**0.288**	**0.279**
HW-TSC	0.165	0.166	0.235	0.266
BASELINE	0.156	0.167	0.219	0.227
KUNMT	–	0.214	0.210	–
NJUNLP	–	**0.284**	–	–

Table 7: F1-score for the official submissions to WMT23 Quality Estimation **Task 2 Error Span Detection**. For each language pair, results marked in bold correspond to the winning submissions, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey

quality tags. This inferred alignment between fine-grained quality annotations and overall quality at the segment level seems to be a promising direction for further improvements in quality estimation.

6.2 Task 2

For Task 2, the submissions are ranked using the F1-score, computed at character level for the annotated error spans, as described in Section 3.2. Precision and Recall scores are also provided as complementary information to help contextualise the performance observed. Statistical significance was computed using randomisation tests (Yeh, 2000) with Bonferroni correction (Abdi, 2007) for each language pair. The results for Task 2 are described in Table 7.

For this subtask we also had participants using pretrained large language models to enhance their submissions. Both KUNMT and Unbabel-IST (for the complementary submission of the latter) used GPT-4 with prompts tailored to fine-grained error span detection. KUNMT use an approach where they combine two prompts in a chain-of-thought manner, asking the model to act as an expert that either:

- Acts as an expert annotator that evaluates the translation and annotates error spans and severities (following the task instructions); or
- Acts as annotation validator and edits previous annotations or marks them as good.

We provide the full prompts in the Appendix E. In turn, Unbabel considers GPT4 both for the word level part of Task 1 and for Task 2, using prompts inspired by (Fernandes et al., 2023).

Aside from the use of LLMs, there are two main approaches in participating submissions: *i*) Participants who extended the word-level approach to obtain fine-grained error spans (HW-TSC, NJUNLP);

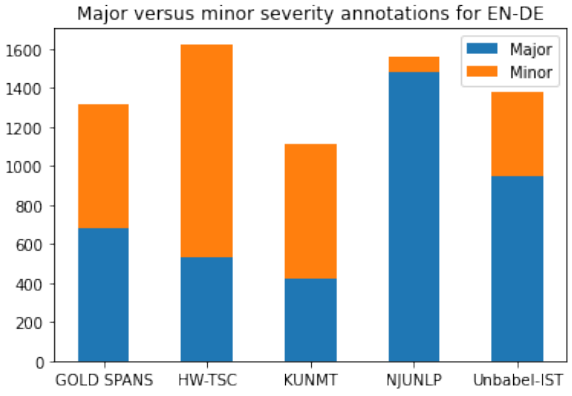


Figure 3: Balance between major and minor errors for each team and the original annotation (first bar).

and *ii*) participants who designed a methodology directly targeting error-spans (KUNMT, Unbabel-IST). To expand the word-level approach, NJUNLP maintain the binary OK/BAD labels and tune different thresholds to decide the severity of the BAD tags. They resolve severity inconsistencies over spans by taking the worst severity. Instead, HW-TSC convert the binary task to a 3-way classification ([OK, MINOR, MAJOR]) and tune on the Task 2 data. On the other hand, in their main approach, Unbabel-IST depart from the word-level approach and tune XCOMET directly on error span annotations, in a multitask setting (predicting both overall sentence score, and severities over spans).

Analysing the results, we can observe that with the exception of KUNMT, most submissions obtain higher *Recall* scores compared to *Precision*. Moreover, if we look at the distribution of identified error severities for e.g. En-De (see Figure 3) we can also observe a difference in the severity proportions as NJUNLP and Unbabel-IST identify mostly "major" errors and obtain better performance compared to KUNMT and HW-TSC that predict less skewed severities, with proportions closer to the gold data.

Best performers. Four teams participated in Task 2, IST-Unbabel, HW-TSC, KUNMT and NJUNLP, with only Unbabel-IST and HW-TSC participating in all tasks. NJUNLP ranked first for the En-De language pair while Unbabel-IST ranked first for the rest, including the multilingual track. We note that both top-ranking participants are using ensemble approaches as well as enhancing their approaches with pseudo-data (pseudo-references in the case of Unbabel-IST and pseudo-MQM scores in the case of NJUNLP).

7 Discussion

In what follows, we discuss the main findings of this year's shared task based on the goals we had previously identified for it.

Granularity of quality annotations We note that while performance for the sentence-level quality scores is better, the finer-grained annotations are also contributing towards these improvements. Indeed, while in comparison participants achieve lower performance for the fine-grained and word-level tasks, **most of the top-ranking submissions across tasks constitute multi-task approaches**, that combined information from the finer-grained annotations and the sentence level scores.

Moreover, comparing the F1-scores between the word-level Task 1 and Task 2, while F1-scores for Task 2 are somewhat lower, considering the additional complexity of considering error severities and multiple spans, the performance seems promising. To further encourage future participation and improved performance we aim to focus on extending MQM annotations for the next iterations, but also revising the error severity definition to potentially include '*critical*' errors.

Zero shot predictions We observe that performance for the zero-shot language pairs, He-En and En-Fa was not hampered by the lack of training and development resources. While fewer participants submitted predictions for these languages, their performance was on par with other language pairs. For for the word-level task, scores were actually higher than those observed for other language pairs. Looking closer at the approach adopted by the participants in these tasks, we can see that besides relying on multi-lingual encoders no additional data was used to train for these languages, across tasks. These findings are encouraging towards annotating a wider range of language pairs (maintaining the emphasis on low and medium source languages) to test on for the upcoming editions, even when training resources are scarce.

Hallucinations We report the hallucination detection results in Table 9. Overall, the results indicate that good-quality QE models are capable of detecting hallucinations very satisfactorily. Some submissions, in fact, obtain perfect or near-perfect results for some language pairs, which indicates that they are able to appropriately penalise the severity of hallucination errors. Not only that, but they can

En-De		Zh-En		He-En		En-Mr		En-Hi		En-Ta		En-Te		En-Gu	
No. Halls	Total	No. Halls	Total	No. Halls	Total	No. Halls	Total	No. Halls	Total	No. Halls	Total	No. Halls	Total	No. Halls	Total
9	1422	13	1260	48	896	86	836	74	824	75	824	75	811	75	825

Table 8: Statistics (number of hallucinations and total number of samples) of the hallucination detection evaluation sets for each language pair.

	En-De		Zh-En		He-En		En-Mr		En-Hi		En-Ta		En-Te		En-Gu	
Model	AUC	R@k	AUC	R@k	AUC	R@k	AUC	R@k	AUC	R@k	AUC	R@k	AUC	R@k	AUC	R@k
NJUNLP	96.25	44.44	–	–	41.62	25.00	–	–	–	–	–	–	–	–	–	–
Unbabel-IST	99.99	88.89	98.97	84.62	97.43	83.33	99.65	91.86	99.92	97.30	99.82	92.00	99.83	93.33	99.82	92.00
IOL Research	100.00	100.00	99.60	84.62	97.04	87.50	98.59	87.21	99.83	97.30	99.48	89.33	99.51	88.00	99.95	98.67
BeringLab	80.11	0.00	90.40	7.69	–	–	–	–	–	–	–	–	–	–	–	–
MMT	90.71	0.00	96.97	30.77	–	–	92.53	62.79	95.85	78.38	91.29	65.33	96.57	68.00	52.58	8.00
HW-TSC	99.97	88.89	98.77	76.92	–	–	99.51	89.53	99.42	95.95	99.49	93.33	99.99	97.33	99.94	98.67
Surrey AI	–	–	–	–	–	–	85.79	48.84	98.86	85.14	98.24	82.67	98.82	82.67	99.28	88.00

Table 9: Hallucination detection performance for the official submissions to **Task 1 Sentence-level**. We report AUROC (↑) and R@k, in percentage (↑).

Source:
大学于1896至1991年，及1912至1968年间分别接管了王家音乐学院和王家安大略博物馆；后两者如今作为独立的机构仍与大学保留着密切的联系。

Hallucination in zh-en not detected by the Unbabel-IST and IOL Research submissions:
The University took over the **Oscillation:Royal Academy of Music and the Royal Academy of Music** between 1896 and 1991, and between 1912 and 1968, and the Royal Academy of Music and the Royal Academy of **Oscillation:Art and Art** Museum, which are now in close connection with the University as an independent institution.

Source:
Intelligent nursing care bed / smart bed Nursing care beds with technical equipment including sensors and notification functions are known as “intelligent” or “smart” beds.

Hallucination in en-de not detected by the Unbabel-IST submission:
Intelligente Pflege-Bett / Smart-**Oscillation: Bett-Bett-Bett-Bett-Bett-Bett-Bett** mit technischer Ausstattung einschließlich Sensoren und Notifizierungsfunktionen sind als "intelligente" oder "intelligente" Bett-Bett bekannt.

Table 10: Examples of hallucinations not flagged in the bottom k of the respective language pair by the best systems in Task 1 Sentence-level.

also distinguish the hallucinations from other translations whose quality may not necessarily be high. This is a property that was not observed in previous iterations of QE models (Amrhein and Sennrich, 2022; Raunak et al., 2022), in particular those that were based on dual encoding of the source and the translation (e.g., COMET-QE (Rei et al., 2020)).

Nevertheless, even the top-performing QE systems, including the winning submissions, may struggle with localised critical errors such as oscillations. We show two such examples in Table 10. In fact, although most pathological hallucinations are detected, some egregious examples have not been detected by both the IOL Research and Unbabel-IST systems (e.g., a he-en translation that contains 70 hallucinated <unk> tokens, and a zh-en translation that contains the oscillation "Tropical and Sub-Tropical Plains and Plains, Tropical and Sub-Tropical Plains and Plains, Tropical and Sub-Tropical High Plains, Sub-Tropical Plains and Sub-Tropical Plains, Sub-Tropical Plains and Sub-Tropical Plains").

One hypothesis for this undesirable behavior is that such samples are out-of-distribution for the QE systems. As such, augmenting the training sets with examples of such hallucinations (e.g., as done in XCOMET (Guerreiro et al., 2023b)) may be a straightforward yet effective approach for correcting this behavior.

8 Conclusions

This year's edition of the QE Shared Task introduced a number of new elements: new low-resource language pairs (including two zero-shot ones), new test sets, and new fine-grained error detection task that we aspire to continue in future editions. It also introduced a mix of hallucinated data together with the original translations, allowing us to assess the robustness of submissions and detect failure patterns that will hopefully help develop more robust QE systems in the future.

The tasks attracted a steady number of participating teams and we believe the overall results are a great reflection of the evolution of the QE field. We note that the gold labels and best submissions to all tasks are made available for those interested in further analysing the results. We aspire for the future editions to continue the efforts set in this and previous years and expand the resources and coverage of QE, while further exploring recent and more challenging subtasks such as fine-grained QE and explainable QE.

9 Ethical Considerations

MQM and DA annotations in this paper are done by professional translators. They are all paid at professional rates.

Organisers from Unbabel and University of Surrey have submitted to this task without using prior access to test sets nor using any insider information.

Acknowledgements

Part of this work was supported by EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), by the project DECOLLAGE (ERC-2022-CoG 101088763), by Fundação para a Ciência e Tecnologia through contract UIDB/50008/2020, and by the Portuguese Recovery and Resilience Plan through project C645008882- 00000055 (Center for Responsible AI).

We thank the annotation agencies Zibanka Media Services Pvt. Ltd. and Techliebe for working with us towards annotating DA data for Indic language pairs. We also thank the European Association for Machine Translation (EAMT) for sponsoring our Indic language pair annotation project at the University of Surrey.

References

Hervé Abdi. 2007. The bonferroni and šidák corrections for multiple comparisons. *Encyclopedia of measurement and statistics*, 3:103–107.

Chantal Amrhein and Rico Sennrich. 2022. Identifying weaknesses in machine translation metrics through minimum Bayes risk decoding: A case study for COMET. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1125–1141, Online only. Association for Computational Linguistics.

Fatemeh Azadi, Heshaam Faili, and Mohammad Javad Dousti. 2022. Mismatching-Aware Unsupervised Translation Quality Estimation for Low-Resource Languages. *arXiv preprint arXiv:2208.00463*.

Aljoscha Burchardt. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Qu Cui, Shujian Huang, Jiahuan Li, Xiang Geng, Zaixiang Zheng, Guoping Huang, and Jiajun Chen. 2021. Directqe: Direct pretraining for machine translation quality estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12719–12727.

David Dale, Elena Voita, Loic Barrault, and Marta R. Costa-jussà. 2023a. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity Even better. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–50, Toronto, Canada. Association for Computational Linguistics.

David Dale, Elena Voita, Janice Lam, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Loïc Barrault, and Marta R. Costa-jussà. 2023b. Halomi: A manually annotated benchmark for multilingual hallucination and omission detection in machine translation.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André FT Martins, Graham Neubig, Ankush Garg, Jonathan H Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. *arXiv preprint arXiv:2308.07286*.

Javier Ferrando, Gerard I. Gállego, and Marta R. Costa-jussà. 2022. Measuring the mixing of contextual information in the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8698–8714, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2022. MLQE-PE: A multilingual quality estimation and post-editing dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4963–4974, Marseille, France. European Language Resources Association.

Erick Fonseca, Lisa Yankovskaya, André FT Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the wmt 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Xiang Geng, Zhejian Lai, Yu Zhang, Shimin Tao, Hao Yang, Jiajun Chen, and Shujian Huang. 2023a. Njunlp's participation for the wmt2023 quality estimation shared task. In *Proceedings of the Eigth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Xiang Geng, Yu Zhang, Jiahuan Li, Shujian Huang, Hao Yang, Shimin Tao, Yimeng Chen, Ning Xie, and Jiajun Chen. 2023b. Denoising pre-training for machine translation quality estimation with curriculum learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12827–12835.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Nuno M. Guerreiro, Duarte Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023a. Hallucinations in large multilingual translation models.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023b. xcomet: Transparent machine translation evaluation through fine-grained error detection.

Nuno M. Guerreiro, Elena Voita, and André Martins. 2023c. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075, Dubrovnik, Croatia. Association for Computational Linguistics.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yuang Li, Chang Su, Ming Zhu, Mengyao Piao, Xinglin Lyu, Min Zhang, and Hao Yang. 2023. Hw-tsc 2023 submission for the quality estimation shared task. In *Proceedings of the Eigth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Arle Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. 2014. Using a new analytic measure for the annotation and analysis of mt errors on real data. In *Proceedings of the 17th Annual conference of the European Association for Machine Translation*, pages 165–172.

Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Tharindu Ranasinghe, Constantin Orašan, and Ruslan Mitkov. 2020. Transquest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5070–5081.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Vikas Raunak, Matt Post, and Arul Menezes. 2022. SALTED: A framework for SAlient long-tail translation error detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5163–5179, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up cometkiwi: Unbabel-ist 2023 submission for the quality estimation shared task. In *Proceedings of the Eigth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José GC de Souza, Taisiya Glushkova, Duarte M Alves, Alon Lavie, et al. 2022a. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. *WMT 2022*, page 634.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 634–645, Abu Dhabi. Association for Computational Linguistics.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia.

Archchana Sindhujan, Diptesh Kanojia, Constantin Orăsan, and Tharindu Ranasinghe. 2023. Surreyai 2023 submission for the quality estimation shared task. In *Proceedings of the Eigth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, Online. Association for Computational Linguistics.

Shimin Tao, Su Chang, Ma Miaomiao, Hao Yang, Xiang Geng, Shujian Huang, Min Zhang, Jiaxin Guo,

Minghan Wang, and Yinglu Li. 2022. Crossqe: Hwtsc 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 646–652.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Dusan Varis and Ondřej Bojar. 2021. Sequence length is a domain: Length-based overfitting in transformer models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8246–8257, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Evan J. Williams. 1959. *Regression Analysis*, volume 14. Wiley, New York, USA.

Yulong Wu, Viktor Schlegel, Daniel Beck, and Riza Batista-Navarro. 2023. Mmt's submission for the wmt 2023 quality estimation shared task. In *Proceedings of the Eigth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Zeyu Yan. 2023. Iol research's submission for wmt 2023 quality estimation shared task. In *Proceedings of the Eigth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Alexander Yeh. 2000. More Accurate Tests for the Statistical Significance of Result Differences. In *Coling-2000: the 18th Conference on Computational Linguistics*, pages 947–953, Saarbrücken, Germany.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

A Hyper-parameters of pre-trained baseline models for Task 1 Quality Estimation

Hyper-parameter	**T1 Sentence-level** COMET-QE	**T1 Word-level** COMETKIWI
Encoder Model	XLM-RoBERTa (large)	XLM-RoBERTa (large)
Optimizer	Adam (default parameters)	Adam (default parameters)
n frozen epochs	0.3	0.3
Keep embeddings frozen	True	True
Learning rate	3e-05 and 1e-05	1e-06 and 1e-05
Batch size	4	4
Loss function	MSE	MSE and CE ($\epsilon = 1.0$)
Dropout	0.15	0.1
FP precision	32	32
Feed-Forward hidden units	[2048, 1024]	[2048, 1024]
Word weights	–	[0.3, 0.7]
Feed-Forward activation	Tanh	–

Table 11: Hyper-parameters of both the COMET-QE and the CometKiwi models used as baselines for Task 1 Quality Estimation.

B Official Results of the WMT23 Quality Estimation Task 1 Sentence-level

Tables 12, 13, 14, 15, 16, 17, 18, 19 and 20 show the results for all language pairs and the multilingual variants, ranking participating systems best to worst using Spearman correlation as primary key for each of these cases.

Model	**Spearman**	Pearson	Kendall
Unbabel-IST •	0.594	0.580	0.438
IOL Research	0.556	0.513	0.407
BASELINE	0.372	0.308	0.265

Table 12: Official results of the WMT23 Quality Estimation Task 1 Sentence-level **Multilingual** (average over all language pairs). Teams marked with "•" are the winners, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey.

Model	**Spearman**	Pearson	Kendall	Disk footprint (B)	# Model params	Ensemble
IOL Research •	0.483	0.429	0.364	2,357,242,105	589,270,071	5
NJUNLP •	0.479	0.423	0.360	3,264,730,349	560,145,557	12
Unbabel-IST	0.456	0.457	0.346	42,868,104,221	10,716,932,147	6
HW-TSC	0.437	0.433	0.331	27,730,527,504	6,932,631,876	12
BeringLab	0.380	0.281	0.283	2,243,955,309	560,945,155	3
BASELINE	0.340	0.253	0.257	2,277,430,715	569,330,715	1
MMT	0.316	0.221	0.237	2,448,132,038	569,330,715	6

Table 13: Official results of the WMT23 Quality Estimation Task 1 Sentence-level for **Englih-German (MQM)**. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey.

Model	**Spearman**	Pearson	Kendall	Disk footprint (B)	# Model params	Ensemble
Unbabel-IST •	0.493	0.423	0.378	42,868,104,221	10,716,932,147	4
IOL Research •	0.482	0.392	0.369	2,357,242,105	589,270,071	5
HW-TSC	0.460	0.369	0.352	27,730,527,504	6,932,631,876	12
BASELINE	0.447	0.318	0.342	2,277,430,715	569,330,715	1
BeringLab	0.384	0.230	0.288	2,243,955,309	560,945,155	3

Table 14: Official results of the WMT23 Quality Estimation Task 1 Sentence-level for **Chinese-English (MQM)**. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey.

Model	**Spearman**	Pearson	Kendall	Disk footprint (B)	# Model params	Ensemble
Unbabel-IST •	0.668	0.518	0.499	42,868,104,221	10,716,932,147	4
IOL Research	0.575	0.424	0.416	2,357,242,105	589,270,071	5
BASELINE	0.475	0.363	0.337	2,277,430,715	569,330,715	1

Table 15: Official results of the WMT23 Quality Estimation Task 1 Sentence-level for **Hebrew-English (MQM)**. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey.

Model	**Spearman**	Pearson	Kendall	Disk footprint (B)	# Model params	Ensemble
Unbabel-IST •	0.704	0.735	0.513	42,868,104,221	10,716,932,147	6
HW-TSC •	0.692	0.718	0.504	27,730,527,504	6,932,631,876	12
MMT	0.650	0.663	0.466	2,448,132,038	569,330,715	7
SurreyAI	0.596	0.668	0.423	2,362,232,012	633,305,686	3
IOL Research	0.505	0.372	0.353	2,357,242,105	589,270,071	5
BASELINE	0.392	0.427	0.274	2,277,430,715	569,330,715	1

Table 16: Official results of the WMT23 Quality Estimation Task 1 Sentence-level **English-Marathi (DA)**. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey.

Model	**Spearman**	Pearson	Kendall	Disk footprint (B)	# Model params	Ensemble
HW-TSC •	0.644	0.720	0.477	27,730,527,504	6,932,631,876	12
IOL Research	0.600	0.667	0.433	2,357,242,105	589,270,071	5
Unbabel-IST	0.598	0.667	0.431	42,868,104,221	10,716,932,147	4
SurreyAI	0.551	0.668	0.395	2,362,232,012	633,305,686	3
MMT	0.494	0.570	0.345	2,448,132,038	569,330,715	7
BASELINE	0.281	0.245	0.190	2,277,430,715	569,330,715	1

Table 17: Official results of the WMT23 Quality Estimation Task 1 Sentence-level **English-Hindi (DA)**. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey.

Model	**Spearman**	Pearson	Kendall	Disk footprint (B)	# Model params	Ensemble
HW-TSC •	0.775	0.778	0.597	27,730,527,504	6,932,631,876	12
IOL Research	0.740	0.742	0.557	2,357,242,105	589,270,071	5
Unbabel-IST	0.739	0.733	0.550	42,868,104,221	10,716,932,147	4
SurreyAI	0.674	0.710	0.495	2,362,232,012	633,305,686	3
MMT	0.547	0.531	0.384	2,448,132,038	569,330,715	7
BASELINE	0.507	0.402	0.354	2,277,430,715	569,330,715	1

Table 18: Official results of the WMT23 Quality Estimation Task 1 Sentence-level **English-Tamil (DA)**. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey.

Model	**Spearman**	Pearson	Kendall	Disk footprint (B)	# Model params	Ensemble
HW-TSC •	0.394	0.350	0.269	27,730,527,504	6,932,631,876	12
Unbabel-IST •	0.388	0.362	0.264	42,868,104,221	10,716,932,147	4
IOL Research	0.376	0.344	0.257	2,357,242,105	589,270,071	5
SurreyAI	0.349	0.376	0.241	2,362,232,012	633,305,686	3
MMT	0.337	0.281	0.228	2,448,132,038	569,330,715	7
BASELINE	0.193	0.153	0.134	2,277,430,715	569,330,715	1

Table 19: Official results of the WMT23 Quality Estimation Task 1 Sentence-level **English-Telegu (DA)**. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey.

Model	**Spearman**	Pearson	Kendall	Disk footprint (B)	# Model params	Ensemble
Unbabel-IST •	0.714	0.745	0.529	42,868,104,221	10,716,932,147	4
IOL Research	0.695	0.742	0.513	2,357,242,105	589,270,071	5
HW-TSC	0.691	0.714	0.511	27,730,527,504	6,932,631,876	12
SurreyAI	0.649	0.700	0.474	2,362,232,012	633,305,686	3
MMT	0.540	0.581	0.386	2,448,132,038	569,330,715	7
BASELINE	0.337	0.307	0.230	2,277,430,715	569,330,715	1

Table 20: Official results of the WMT23 Quality Estimation Task 1 Sentence-level **English-Gujarati (DA)**. Teams marked with "•" are the winners, as they are not significantly outperformed by any other system according to the Williams Significance Test (Williams, 1959). Baseline systems are highlighted in grey.

C Official Results of the WMT23 Quality Estimation Task 1 Word-level

Tables 21, 22, 23, 24, 25 and 26 show the results for all language pairs and the multilingual variant, ranking participating systems best to worst using Matthews Correlation Coefficient (MCC) as primary key for each of these cases.

Model	**MCC**	F1-score
Unbabel-IST •	0.329	0.355
IOL Research	0.298	0.322
BASELINE	0.252	0.243

Table 21: Official results of the WMT23 Quality Estimation Task 1 Word-level **Multilingual** (average over all language pairs). The winning submission is indicated by a •. Baseline systems are highlighted in grey.

Model	**MCC**	F1-score	Disk footprint (B)	# Model params	Ensemble
NJUNLP •	0.297	0.329	3,264,730,349	560,145,557	12
IOL Research	0.256	0.281	2,357,242,105	589,270,071	5
Unbabel-IST	0.246	0.279	2,252,351,787	563,041,309	1
BeringLab	0.233	0.269	2,243,955,309	560,945,155	5
BASELINE	0.179	0.207	2,252,351,659	563,041,309	1

Table 22: Official results of the WMT23 Quality Estimation Task 1 Word-level **English-German (MQM)**. The winning submission is indicated by a •. Baseline systems are highlighted in grey.

Model	**MCC**	F1-score	Disk footprint (B)	# Model params	Ensemble
Unbabel-IST •	0.302	0.331	2,252,351,787	563,041,309	1
IOL Research	0.250	0.268	2,357,242,105	589,270,071	5
BeringLab	0.241	0.262	2,243,955,309	560,945,155	5
BASELINE	0.225	0.255	2,252,351,659	563,041,309	1

Table 23: Official results of the WMT23 Quality Estimation Task 1 Word-level **Chinese-English (MQM)**. The winning submission is indicated by a •. Baseline systems are highlighted in grey.

Model	**MCC**	F1-score	Disk footprint (B)	# Model params	Ensemble
Unbabel-IST •	0.402	0.439	2,252,351,787	563,041,309	1
IOL Research	0.359	0.361	2,357,242,105	589,270,071	5
BASELINE	0.275	0.275	2,252,351,659	563,041,309	1

Table 24: Official results of the WMT23 Quality Estimation Task 1 Word-level **Hebrew-English (MQM)**. The winning submission is indicated by a •. Baseline systems are highlighted in grey.

Model	**MCC**	F1-score	Disk footprint (B)	# Model params	Ensemble
Unbabel-IST •	0.347	0.359	2,252,351,787	563,041,309	1
IOL Research	0.334	0.373	2,357,242,105	589,270,071	5
BASELINE	0.287	0.224	2,252,351,659	563,041,309	1

Table 25: Official results of the WMT23 Quality Estimation Task 1 Word-level **English-Marathi (Post-Editing)**. The winning submission is indicated by a •. Baseline systems are highlighted in grey.

Model	**MCC**	F1-score	Disk footprint (B)	# Model params	Ensemble
IOL Research •	0.351	0.389	2,357,242,105	589,270,071	5
Unbabel-IST	0.345	0.365	2,252,351,787	563,041,309	1
BASELINE	0.293	0.254	2,252,351,659	563,041,309	1

Table 26: Official results of the WMT23 Quality Estimation Task 1 Word-level **English-Farsi (Post-Editing)**. The winning submission is indicated by a •. Baseline systems are highlighted in grey.

D Official Results of the WMT23 Quality Estimation Task 2 Error Span Detection

Tables 27, 28, 29 and 30 show the results for all language pairs and the multilingual variant, ranking participating systems best to worst using F1-score as primary key for each of these cases.

Model	**F1-score**	Precision	Recall
Unbabel-IST •	0.220	0.164	0.360
HW-TSC	0.165	0.177	0.161
BASELINE	0.156	0.203	0.128

Table 27: Official results of the WMT23 Quality Estimation Task 1 Word-level **Multilingual** (average over all language pairs). The winning submission is indicated by a •. Baseline systems are highlighted in grey.

Model	**F1-score**	Precision	Recall	Disk footprint (B)	# Model params	Ensemble
NJUNLP •	0.284	0.238	0.352	560,145,557	3,264,730,349	12
Unbabel-IST	0.273	0.209	0.394	-1	-1	-1
KUNMT	0.214	0.224	0.206	818,245,780	2,235,540,305	3
BASELINE	0.167	0.229	0.131	563,041,309	2,252,351,659	1
HW-TSC	0.166	0.220	0.133	285,019,112	1,148,646,407	5

Table 28: Official results of the WMT23 Quality Estimation Task 1 Word-level **English-German (MQM)**. The winning submission is indicated by a •. Baseline systems are highlighted in grey.

Model	**F1-score**	Precision	Recall	Disk footprint (B)	# Model params	Ensemble
Unbabel-IST •	0.288	0.246	0.349	3,485,770,281	13,943,358,015	5
HW-TSC	0.235	0.221	0.250	285,019,112	1,148,646,407	4
BASELINE	0.219	0.259	0.190	563,041,309	2,252,351,659	1
KUNMT	0.210	0.216	0.204	818,245,780	2,235,540,305	3

Table 29: Official results of the WMT23 Quality Estimation Task 1 Word-level **Chinese-English (MQM)**. The winning submission is indicated by a •. Baseline systems are highlighted in grey.

Model	**F1-score**	Precision	Recall	Disk footprint (B)	# Model params	Ensemble
Unbabel-IST •	0.279	0.241	0.332	3,485,770,281	13,943,358,015	5
HW-TSC	0.266	0.254	0.279	285,019,112	1,148,646,407	10
BASELINE	0.227	0.474	0.150	563,041,309	2,252,351,659	1

Table 30: Official results of the WMT23 Quality Estimation Task 1 Word-level **Hebrew-English (MQM)**. The winning submission is indicated by a •. Baseline systems are highlighted in grey.

E GPT-4 prompts for Task 2

We add below the prompts used by KUNMT team with GPT4 for Task 2.

Expert annotator prompt:

You are an expert in the Fine-grained error span detection task. The goal of this task is to predict the word-level translation error spans. you will be asked to predict both the error span (start and end indices) as well as the error severity (major or minor) for each segment. There can be multiple error spans, and you must indicate the severity of the error for the spans that exist. If no errors exist in the translation, the error span is (-1,-1) and the error severity is no-error.

Expert validator prompt:

Review this result by checking the work done by the other workers. If the work was done correctly, mark it as "GOOD"; if there were any errors, re-annotate the Error Span and Error Severity. To avoid

inconsistencies, we expect the indices of the errors spans to correspond to characters in the target string before tokenisation, i.e., the target string that will be provided as test data.'

Findings of the Word-Level AutoCompletion Shared Task in WMT 2023*

Lemao Liu [1] **Francisco Casacuberta** [2] **George Foster** [3] **Guoping Huang** [1]
Philipp Koehn [4] **Geza Kovacs** [5] **Shuming Shi** [1] **Taro Watanabe** [6] **Chengqing Zong** [7]
[1] Tencent AI Lab [2] Universitat Politècnica de València [3] Google
[4] Johns Hopkins University [5] LILT [6] Nara Institute of Science and Technology
[7] Institute of Automation, Chinese Academy of Sciences

Abstract

This paper presents the overview of the second Word-Level autocompletion (WLAC) shared task for computer-aided translation, which aims to automatically complete a target word given a translation context including a human typed character sequence. We largely adhere to the settings of the previous round of the shared task, but with two main differences: 1) The typed character sequence is obtained from the typing process of human translators to demonstrate system performance under real-world scenarios when preparing some type of testing examples; 2) We conduct a thorough analysis on the results of the submitted systems from three perspectives. From the experimental results, we observe that translation tasks are helpful to improve the performance of WLAC models. Additionally, our further analysis shows that the semantic error accounts for a significant portion of all errors, and thus it would be promising to take this type of errors into account in future.

1 Introduction

Computer-aided translation (CAT) helps human translators produce high-quality translations with the assistance of machine translation systems (Koehn et al., 2003; Vaswani et al., 2017), and it has witnessed a lot of attention during the past decades (Bowker, 2002; Koehn, 2009; Foster et al., 1997; Langlais et al., 2000; Barrachina et al., 2009; Alabau et al., 2014; Knowles and Koehn, 2016; Santy et al., 2019; Huang et al., 2021). Among all the tasks in CAT, Word-Level autocompletion (WLAC) is one of the most fundamental tasks and its goal is to autocomplete a word when a human translator types a sequence of characters (Huang et al., 2015; Li et al., 2021), in order to accelerate the editing process for human translators under CAT settings. To facilitate the research in WLAC, the first WLAC shared task was held in WMT 2022 (Casacuberta et al., 2022; Yang et al., 2022; Ángel Navarro et al., 2022; Moslem et al., 2022; Ailem et al., 2022). This year, we continue holding the second edition of WLAC shared task in WMT 2023.

In this paper, we summarize the overview of the WLAC shared task in WMT 2023, which is named by WLAC 2023 for brevity, including data preparation process, submitted systems and their evaluation results. Specifically, WLAC 2023 involves two language pairs, i.e. Chinese-English and German-English, and contains four directional sub-tasks in total, similar to WLAC 2022 shared task. For training data preparation, we follow the common practice of leveraging a bilingual corpus for simulation. For test data preparation, however, there is one important difference in this year to make the test data more similar to realistic scenarios: for some testing examples (see §2.2), their typed character sequences are obtained from the typing process of human translators.

We have received twenty-one submissions in total from four teams in WLAC 2023. We evaluate all these submissions and present their overall evaluation results. In particular, we conduct a thorough analysis of submitted systems to better understand the challenges and difficulties emerged in WLAC tasks. The analysis of these systems is investigated according to three perspectives which include the frequency of target words, the size of context, as well as the human defined error types. From all the perspectives, we observe some insights which might be useful for further improvement on WLAC in future. In summary, our main findings are highlighted as follows:

1. Through effective use of translation models, it is able to substantially benefit the WLAC models in terms of accuracy.

2. Among all type of errors, the semantic error

* The authors are listed alphabetically except the first author.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 654–662
December 6–7, 2023.

makes up the majority of error cases, where predicted words are semantically deviated from the ground-truth words.

3. It is possible to directly use large language models (LLMs) for WLAC tasks, but the results show that currently LLMs can not effectively handle WLAC without fine-tuning.

2 Task Description and Data Preparation

2.1 Task Definition

WLAC tasks aim to auto-complete a target word for the CAT process. The definition of WLAC is as follows: given a source sequence x, translation context $c = (c_l, c_r)$, where c_l and c_r are left and right side context respectively, and a character-level typed sequence s by human translators, WLAC aims to predict the target word w with s as its prefix, which should be the most appropriate to be placed between c_l and c_r (Huang et al., 2015; Li et al., 2021). Formally, we expect to model the relationship following the below equation:

$$w = f(x, s, c_l, c_r) \quad (1)$$

More generally, the right or left side context could be empty in real-world CAT systems. Consequently, there are four types of situations should be considered in WLAC tasks:

1. zero-context: both c_l and c_r are empty;
2. suffix: c_l is empty while c_r is non empty;
3. prefix: c_l is non empty while c_r is empty;
4. bi-context: both c_l and c_r are non empty.

	EN-DE	EN-ZH
Sentence Pairs	4,465,840	15,886,041
Words (src/tgt)	120M/114M	441M/395M

Table 1: The statistical description of the total number of sentence pairs and the scale of tokenized words on English ⇔ German and English ⇔ Chinese language pairs.

2.2 Data Preparation

We mainly follow the previous edition settings for data preparation, which includes two language pairs, i.e. English ⇔ Chinese and English ⇔ German. Both translation directions are considered in the evaluation, resulting in four directional tasks.

Training Data Following previous edition settings, we employ simulated training data $\langle x, s, c, w \rangle$ for this year WLAC. The construction of which follows the algorithm proposed by Li et al. (2021) [1]. The reason of such a simulation is to compensate for the limited size of manually annotated training data.

Specifically, for English ⇔ German language pair, we use the WMT14 EN-DE training dataset preprocessed by Stanford NLP Group [2], which is about 4.5 million sentence pairs; For English ⇔ Chinese pair, we leverage UN Parallel Corpus dataset [3] from WMT17, which consists of 15 million sentence pairs. Moses tokenizer [4] is applied to both English and German sentences while Jieba [5] is used to segment Chinese sentences. The detailed statistical description of the datasets is shown in Table 1.

For a fair comparison, only the above-mentioned corpus is allowed to be employed for bilingual training. However, there is no limitation for any monolingual data usage and even for pre-trained language models (Devlin et al., 2018) or large language models such as ChatGPT and Llama (Touvron et al., 2023).

Testing Data Similar to the data preparation in WLAC 2022, testing data in this year consists of two types of datasets as well. **Type I** is the conventional simulation on bilingual data which follows the same construction rules as the training data; **Type II** testing data is obtained from the real-world post-editing scenario. To alleviate any information leakage about the testing sets, the bilingual dataset and post-editing data are created by a third-party company [6] to guarantee that both data are not included in the training data.

In details, to create the testing examples for Type II testing set, we focus on the words that the translators had modified and then sample their context according to four types. [7] In particular, unlike WLAC 2022 where the typed sequence is randomly sam-

[1] The scripts for simulation is available at https://github.com/lemaoliu/WLAC.

[2] https://nlp.stanford.edu/projects/nmt/data

[3] https://conferences.unite.un.org/UNCorpus/Home/DownloadOverview

[4] https://github.com/moses-smt/mosesdecoder

[5] https://github.com/fxsjy/jieba

[6] We paid about 10,000 dollars to obtain the test data from the third-party company.

[7] Since the sentences from post-editing naturally belong to bi-context type, we need to obtain all types of examples via randomly sampling context.

Date Type	ZH⇒EN	EN⇒ZH	DE⇒EN	EN⇒DE
Sentence Pairs				
Type I	11341	11430	9653	9367
Type II	5044	5173	4910	5172
Overall	16385	16603	14539	14564
Averaged Length (src/tgt)				
Type I	28.88/4.71	31.88/4.46	28.73/4.47	29.18/4.43
Type II	32.29/5.71	35.66/5.42	32.93/5.46	33.61/5.24
Overall	29.16/4.85	32.22/4.58	29.19/4.59	29.72/4.53

Table 2: The total number of testing examples for both Type I and II cases over four language pair directions. A/B denotes that A is the averaged number of source words in the source sentences and B is the averaged number of target words in the context.

Data Type	ZH⇒EN	EN⇒ZH	DE⇒EN	EN⇒DE
Bi-context				
Type I	2489	2514	2081	2021
Type II	1107	1139	1060	1117
Overall	3596	3653	3141	3138
Prefix				
Type I	3884	3902	3416	3315
Type II	1729	1766	1739	1830
Overall	5613	5668	5155	5145
Suffix				
Type I	2499	2534	2098	2033
Type II	1113	1147	1066	1123
Overall	3612	3681	3164	3156
Zero-Context				
Type I	2466	2479	2058	1997
Type II	1098	1122	1046	1103
Overall	3564	3601	3104	3100

Table 3: The number of testing examples on four types of context cases for each sub-tasks.

pled according to target words, in WLAC 2023 the typed sequences for Type II dataset are obtained according to the typing process of human translators. This makes examples in Type II data more realistic than those in WLAC 2022.

Finally, when generating testing examples from the parallel sentences and post-edited sentences, we increase the proportion of *Prefix* type this year because the *Prefix* context type is more likely to match the popular left-to-right interactive translation systems. The statistics of sentence pairs are shown on Table 2 and the statistics of the different context types are shown on Table 3.

3 Experimental Setting

3.1 Evaluation Metric

According to the findings from WLAC 2022 (Casacuberta et al., 2022), the automatic evaluation result is highly consistent with the human evaluation result on the same dataset. Hence, in this year, we only employ the automatic evaluation for the submitted systems. Specifically, we use accuracy as the automatic evaluation metric (Li et al., 2021) to demonstrate the performance of all submitted systems:

$$acc = \frac{N_{\text{match}}}{N} \quad (2)$$

where N_{match} is the total number of correctly predicted words and N is the total number of all testing samples.

3.2 Submitted Systems

We received 21 submissions from 4 teams. We briefly summarize their approaches below.

SJTU-MTLAB The SJTU-MTLAB participates in all language directions. They submitted both word-level model and BPE-level model and their BPE-level model performs better (Chen and Wang, 2023). The BPE-level model is based on the Transformer architecture with encoder and decoder, where the encoder take the source sentence and all context as input and the decoder is responsible for generating the target word. They also introduce another decoder to generate the full target sentence, and jointly train the full model with WLAC task and machine translation task. The translation decoder is discarded during inference to maintain a reasonable inference cost. For more details about this system, it can be found in Chen et al. (2023).

Systems	ZH-EN	EN-ZH	EN-DE	DE-EN
Traditional Supervised Method				
SJTU-MTLAB	56.93	61.16	67.27	68.16
HW-TSC	56.40	57.80	66.42	68.10
PRHLT/sys1	-	-	37.05	39.98
PRHLT/sys2	-	-	37.38	43.56
Few-Shot Method				
KnowComp/0-shot	9.82	-	9.72	7.53
KnowComp/1-shot	21.43	-	14.96	15.34
KnowComp/5-shot	27.74	-	21.98	22.95

Table 4: Official evaluation results for all submitted systems. The score is reported in accuracy.

HW-TSC The Huawei Translation Services Center (HW-TSC) participates in all language directions. They model the WLAC task in the BPE level and iteratively generates a subword to compose the prediction word. [8] Specifically, they employ an encoder-decoder architecture, where the encoder encodes the source sentence and the decoder takes as input the target side context. They first train a machine translation task as a baseline and then they fine-tune the baseline with WLAC data and BERT-style MLM data to get the final model.

KnowComp KnowComp group proposes a large language model (LLM) based system for this year's WLAC task. They first randomly sample in-context examples as prompts to obtain the row ChatGPT outputs and extract the final prediction by post-processing (Wu et al., 2023). Specifically, they provide the source sentence x and target sentence with a special token `[mask]` as a placeholder for x (i.e., $(c_l, \texttt{[mask]}, c_r)$), and let LLMs predict the word that should fill in the mask position. Since more than one word may be generated, they search for the first word that starts with the pre-typed sequence s as the final prediction. They evaluate the submitted systems in Chinese ⇒ English, German ⇒ English, and English ⇒ German directions.

PRHLT PRHLT group participates in English ⇔ German and German ⇔ English categories. Their submitted system is developed on a segment-based interactive machine translation (IMT) system (Ángel Navarro et al., 2023). It predicts the results by word correction task based on a sequence of segmented contexts. Moreover, to further enhance the system performance under zero-context situations, they developed a dictionary-based translation module for zero-context word completion. Additionally, they made a second submission which fine-tunes an LLM (mT5) (Xue et al., 2020) to adapt it to the WLAC task. To perform this fine-tuning they created a new parallel dataset in which source sentences are the concatenation of the original source sentences + left context + right context + typed sequence, and the target sentences are the autocompletions.

[8] HW-TSC team does not submit the system report this year, but it is told that the system is very similar to that used in WLAC 2022 by personal communication with the team members.

4 Experimental Result and Analysis

4.1 Evaluation Results

Overall result The overall evaluation results of all submissions are reported on Table 4. The performance of HW-TSC and SJTU-MTLAB are comparable, and both systems perform the best among all the submissions. Both HW-TSC and SJTU-MTLAB make use of the knowledge from machine translation, and the large gains indicate the effectiveness to incorporate WLAC task with machine translation task according to the experiments in the system report of Chen and Wang (2023). Furthermore, we can see that fine-tuning the large mT5 model (PRHLT/sys2) delivers substantial improvements over PRHLT/sys1. It is worth noting that the KnowCamp system does not involve re-training for WLAC tasks, and thereby it is unfair to compare it with other systems which are trained with the large scale of the supervised training data. Anyway, its evaluation result still shows that the large language model can not handle the WLAC task well without fine-tuning on the training data.

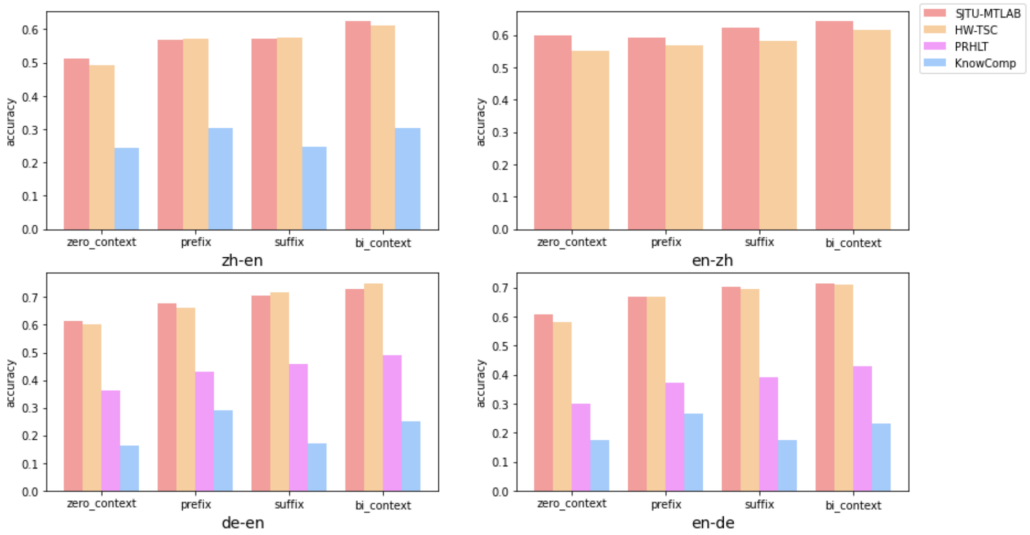


Figure 1: The accuracy of all language directions among different context types.

Result for context types In addition to the overall result, we also evaluate the submissions according to different context types of testing examples for all sub-tasks. The accuracy of four systems for four context types is illustrated in Figure 1, where only the best system from each team is evaluated. As we can see, for most systems, the accuracy increases from *zero_context* to *bi_context*. This indicates that more context can bring better performance. One exception is KnowComp, which does not perform well in *zero_context* and *suffix*. One of the possible reasons is that the large language models would find it difficult to make a correct prediction with little(*zero_context*) or unusual context type (the setting of *suffix* is contradictory with the left-to-right paradigm in large language models).

4.2 Analysis

In this subsection, we conduct a thorough analysis on the evaluation results from three perspectives. Since the analysis results are similar across different language directions, we conduct the following analysis on the de-en direction, because all the systems have submitted results on this direction.

Frequency The first perspective is to analyze the accuracy according to the word frequency. To this end, we divide testing examples into 16 bins according to the frequency of their ground-truth word: suppose an example is with a frequency of f ($f \geq 1$), then it is placed into the bin with id as the rounding number of $\min(16, \log f)$. Then we calculate the accuracy for each bin and the result is depicted in Figure 2. From the figure it is observed that it is very difficult to predict the rare words (i.e. their frequency is zero) in WLAC, which is in line with the task of neural machine translation (Luong et al., 2015). When the frequency is more than one, the accuracy is much higher than that for frequency of zero; however, the accuracy does not strictly increase as the frequency gets higher than one.

Context size The second perspective is to analyze the accuracy of each system according to the context size. The number of words in the left and right contexts indicate whether the context provides the sufficient information to predict the target word and thus the accuracy of each system might be influenced by the context size. Since different examples may have different length in the sentence, we group the examples into bins according to the relative context size defined by the ratio of the context size to the size of the source sentence. Then we measure the accuracy for each bin and the results for all systems are illustrated in Figure 3. As shown from this figure, for all the supervised systems the accuracy generally increases when the relative context size becomes larger. However, the KnowComp system seems to be insensitive to the context size. This fact may indicate that KnowComp does not make full use the context, which provides an explanation why KnowComp does not work well for WLAC.

Error Analysis In order to look deeper into the reason why the models make wrong prediction, we propose to manually analyze the errors made by each system, which is the third perspective. To this end, we first define three types of errors for each

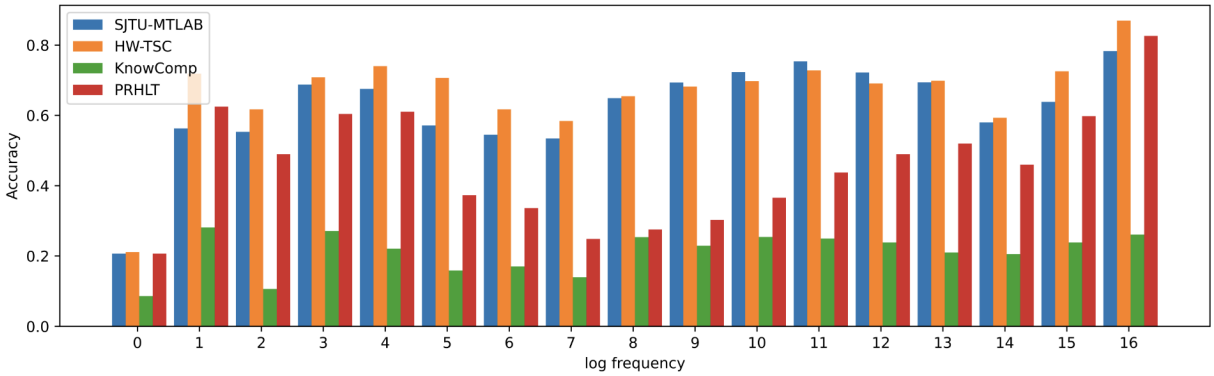


Figure 2: The accuracy of different bins organized according to the frequency of ground-truth target word. Each bin id corresponds to the rounding number of $\min(16, \log f)$ with f as the frequency of the target word.

incorret prediction. The first one is the *constraint error*, where the predicted word fails to meet the constraint of typed character sequence. There are two main reasons to this error: 1) the system does not use the hard constraint manner during inference; 2) the system uses the hard constraint during inferenct but still can not predict a word which satisfies the constraint due to some unusual typed character sequence. Another common type of errors is morphology error, where the prediction has similar semantics with ground truth but has different morphology. For example, the ground truth is *needs* while the prediction is *need*. We detect this type of error by *nltk.stem* [9] tool. The third error is called semantic error, where the predictions are completely deviated from the ground-truth words in semantic. To measure how much the prediction deviate from the ground truth, we use the fastText[10] tool to compute the semantic similarity of predictions and ground truths. We report the proportion of errors where the semantic similarity of prediction and label is less than 0.3.

The results for all systems are reported in Table 5. According to the constraint errors, the SJTU-MTLAB and HW-TSC can meet the constraint well, while the LLM based method, KnowComp, often fails to generate a proper word with given typed character sequence. After manually checking the results from all these systems, we find that both SJTU-MTLAB, PRHLT and KnowComp does not employ the hard constraint during inference and HW-TSC sometimes can not predict a word satisfying the constraint due to unusual typed sequences. In addition, according to the morphology error, as we can see from Table 5, there are still a non-negligible amount of predictions fall into this group, indicating the potential for further improvement. Finally, according to the semantic error, as reported in Table 5, most of the errors belongs to this group. This is the most critical error type, and we recommend reducing this part of the error is very promising to improve the overall performance.

4.3 Discussion on future direction

Through the overall results and analysis, we point out some possible direction of further improvement:

- Incorporating machine translation task. The SJTU-MTLAB and HW-TSC introduce machine translation into the WLAC task and show superior performance. This indicates the importance of adding translation knowledge into WLAC and we encourage more effective method to combine these two tasks.

- Improving large language models for WLAC. KnowComp employs the large language models through in-context learning for WLAC. Although its performance is not as good as other systems, it still exhibits potential because it does not leverage the large-scale supervised data for training. Indeed, simply fine-tuning the large mT5 model on the supervised data yields respectful results (see PRHLT/sys2). Therefore, it is promising to further improve LLMs by using of the supervised data.

- Alleviating the semantic error. The large amount of semantic error indicates that the current systems still fail to model the problem in many cases. We expect the development of more powerful models to push the SOTA forward by taking semantic error into account.

[9] https://www.nltk.org/api/nltk.stem.html
[10] https://fasttext.cc/

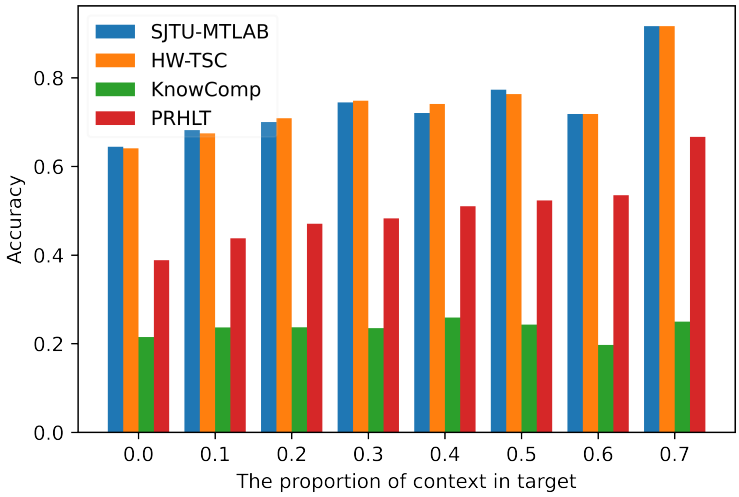


Figure 3: The accuracy of different bins organized according to the relative context size of each example. The relative context size is defined by the ratio of the size of left and right contexts to the size of the source side for each example.

Systems	Constraint	Morphology	Semantic
SJTU-MTLAB	4.12%	12.62%	57.69%
HW-TSC	1.89%	10.98%	56.28%
KnowComp	22.30%	6.77%	74.57%
PRHLT	7.18%	10.23%	59.85%

Table 5: The proportion of different types of error among constraint error, morphology error, and the semantic error respectively. The sum of each line does not equal to 1 because different types or error may share overlaps.

5 Conclusion

This paper presents the overview for the shared task of Word-level Auto-Completion, which is the key component of computer-aided translation. We describe the task definition, data preparation process, the submitted systems, evaluation metric and evaluation results of the systems. We have received twenty-one submissions from four participants this year. We report the evaluation results of all systems, conduct a thorough analysis on the prediction results of these systems and obtain some insightful findings. We hope that our findings can encourage the emerge of more powerful models and attract more researchers to participate the study of computer-aided translation.

Acknowledgements

We would like to appreciate the annotators for their creating test data on this shared task. In addition, we thank the participants for their contributions on this shared task.

References

Melissa Ailem, Jinghsu Liu, Jean-Gabriel Barthélemy, and Raheel Qader. 2022. Lingua custodia's participation at the wmt 2022 word-level auto-completion shared task. In *Proceedings of the Seventh Conference on Machine Translation Shared Tasks Papers*. Association for Computational Linguistics.

Vicent Alabau, Christian Buck, Michael Carl, Francisco Casacuberta, Mercedes García-Martínez, Ulrich Germann, Jesús González-Rubio, Robin L Hill, Philipp Koehn, Luis A Leiva, et al. 2014. Casmacat: A computer-assisted translation workbench. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 25–28.

Sergio Barrachina, Oliver Bender, Francisco Casacuberta, Jorge Civera, Elsa Cubel, Shahram Khadivi, Antonio Lagarda, Hermann Ney, Jesús Tomás, Enrique Vidal, et al. 2009. Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28.

Lynne Bowker. 2002. *Computer-aided translation technology: A practical introduction*. University of Ottawa Press.

Francisco Casacuberta, George Foster, Guoping Huang, Philipp Koehn, Geza Kovacs, Lemao Liu, Shuming Shi, Taro Watanabe, and Chengqing Zong. 2022. Findings of the word-level autocompletion shared task in wmt 2022. In *Proceedings of the Seventh*

Conference on Machine Translation (WMT), pages 812–820.

Xingyu Chen, Lemao Liu, Guoping Huang, Zhirui Zhang, Mingming Yang, Shuming Shi, and Rui Wang. 2023. Rethinking word-level auto-completion in computer-aided translation. In *Proceedings of EMNLP*. Association for Computational Linguistics.

Xingyu Chen and Rui Wang. 2023. Sjtu-mtlab's submission to the wmt23 word-level auto completion task. In *Proceedings of the Seventh Conference on Machine Translation Shared Tasks Papers*. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

George Foster, Pierre Isabelle, and Pierre Plamondon. 1997. Target-text mediated interactive machine translation. *Machine Translation*, 12:175–194.

Guoping Huang, Lemao Liu, Xing Wang, Longyue Wang, Huayang Li, Zhaopeng Tu, Chengyan Huang, and Shuming Shi. 2021. Transmart: A practical interactive machine translation system. *arXiv preprint arXiv:2105.13072*.

Guoping Huang, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2015. A new input method for human translators: integrating machine translation effectively and imperceptibly. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Rebecca Knowles and Philipp Koehn. 2016. Neural interactive translation prediction. In *Conferences of the Association for Machine Translation in the Americas: MT Researchers' Track*, pages 107–120.

Philipp Koehn. 2009. A process study of computer-aided translation. *Machine Translation*, 23:241–263.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

Philippe Langlais, George Foster, and Guy Lapalme. 2000. Transtype: a computer-aided translation typing system. In *ANLP-NAACL 2000 Workshop: Embedded Machine Translation Systems*.

Huayang Li, Lemao Liu, Guoping Huang, and Shuming Shi. 2021. Gwlan: General word-level autocompletion for computer-aided translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4792–4802.

Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 11–19, Beijing, China. Association for Computational Linguistics.

Yasmin Moslem, Rejwanul Haque, and Andy Way. 2022. Word-level auto-completion: What can we achieve out of the box? In *Proceedings of the Seventh Conference on Machine Translation Shared Tasks Papers*. Association for Computational Linguistics.

Sebastin Santy, Sandipan Dandapat, Monojit Choudhury, and Kalika Bali. 2019. Inmt: Interactive neural machine translation prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 103–108.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yi Wu, Haochen Shi, Weiqi Wang, and Yangqiu Song. 2023. Knowcomp submission for wmt23 word-level autocompletion task. In *Proceedings of the Seventh Conference on Machine Translation Shared Tasks Papers*. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Hao Yang, Hengchao Shang, Zongyao Li, Daimeng Wei, Xianghui He, Xiaoyu Chen, Zhengzhe Yu, Jiaxin Guo, Jinlong Yang, Shaojun Li, Yuanchang Luo, Yuhao Xie, Lizhi Lei, and Ying Qin. 2022. Hw-tsc's submissions to the wmt22 word-level auto completion task. In *Proceedings of the Seventh Conference on Machine Translation Shared Tasks Papers*. Association for Computational Linguistics.

Ángel Navarro, Miguel Domingo, and Francisco Casacuberta. 2022. Prhlt's submission to wlac 2022. In *Proceedings of the Seventh Conference on Machine Translation Shared Tasks Papers*. Association for Computational Linguistics.

Ángel Navarro, Miguel Domingo, and Francisco Casacuberta. 2023. Prhlt's submission to wlac 2023. In *Proceedings of the Seventh Conference on Machine Translation Shared Tasks Papers*. Association for Computational Linguistics.

Findings of the WMT 2023 Shared Task on Machine Translation with Terminologies

Kirill Semenov[C] **Vilém Zouhar**[E] **Tom Kocmi**[M] **Dongdong Zhang**[M]
Wangchunshu Zhou[A] **Yuchen Eleanor Jiang**[A]

[C]**Charles University** [E]**ETH Zürich** [M]**Microsoft** [A]**AIWaves**

Abstract

The WMT 2023 Terminology Shared Task investigates progress in machine translation of texts with specialized vocabulary. The participants were given the source text and segment-level terminology dictionaries for three language pairs: Chinese→English, English→Czech, and German→English. We evaluate 21 submissions from 7 teams on two main criteria: general translation quality and the effectiveness of translating specialized terminology. Systems took varied approaches — incorporating terminology at inference time or weakly supervised training that uses terminology access. While incorporating terminology dictionaries leads to improvement in the translation quality, incorporating an equal amount of information from the reference leads to similar results. This challenges the position of terminologies being the crux of meaning in translation, it can also be explained by inadequate metrics which are not terminology-centric.

1 Introduction

General-purpose machine translation models often show limitations when applied to specialized tasks, like translating specialized vocabulary. This gap is critical in medicine, science, and law, where language precision is paramount — medical inaccuracies, juridical misunderstandings, and technological malfunctions can lead to serious problems. The translation of technical terms is not a mere exercise in lexical fidelity — it supports effective communication in highly specialized fields. Terminology correctness and consistency has already been long in focus from the modelling (Dinu et al., 2019; Hasler et al., 2018), evaluation (Zouhar et al., 2020; ibn Alam et al., 2021; Semenov and Bojar, 2022) and translators' perspective (Cabré, 2010; Vargas-Sierra, 2011; Arcan et al., 2017).

We shed light into recent advancement in this area by assessing MT systems with segment-level terminology dictionaries. Alongside the general evaluation of translation quality, our shared task emphasizes the *effectiveness* terminology dictionaries. This task follows the latest efforts on evaluating progress in terminology-enhanced translation (Alam et al., 2021). While we are also concerned with the quality of the translation, we refocus on measuring the relative improvement of incorporating the terminology dictionary. Focusing on System A being overall better with terminologies than System B might obscure the fact that System A is already good without terminologies while the methods of System B improves. From research perspective, System B gives us more insight into how to more efficiently incorporate terminology dictionaries. Additionally, it disentangles the terminology-incorporation methods from the general MT methods.

Source	Der Bericht entspricht FOG.
Reference	The report is ROA-compliant.
Hyp. 1	The report is in accordance with FOG.
Hint 1	"FOG" → "ROA"
Hyp. 2	The report is in accordance with ROA.
Hint 2	"entspricht" → "compliant"
Hyp. 3	The report is compliant with ROA.

Table 1: Translation with "terminologies". Hyp. 1 is without any hints and the worst while Hyp. 3 is close to the reference. Hint 1 is proper terminology while Hint 2 only helps align the translation with the reference. Does terminology-assisted MT work because of Hint 1 or because it leaks information from the reference?

Perf. ↑	**A**	**B**
Base	95	90
+Dict.	92	70

This shared task provides one repackaged and two newly-annotated datasets which can be used for segment-level terminology enhanced machine translation evaluation.[1]

[1]Public terminology datasets Chinese→English (repack), English→Czech (new data) and German→English (new data): github.com/wmt-terminology-task/data-2023

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 663–671
December 6–7, 2023.

German→ English	**Source**: "Most informative is the analysis of airway secretions:" **Reference**: "Häufig jedoch führt die Analyse von Material aus den Atemwegen zur Diagnose:" **Proper**: "analysis of airway secretions" →"Analyse von Material aus den Atemwegen" **Random**: "Most"→"Häufig"
English→ Czech	**Source**: "We present Eman, an experiment manager, and show how to use it to train several simple MT systems." **Reference**: "Popisujeme Emana, nástroj pro správu experimentů, a ukazujeme, jak ho lze využít k trénování několika jednoduchých systémů pro strojový překlad." **Proper**: "Eman"→"Emana", "an experiment manager"→"nástroj pro správu experimentů", "MT systems"→"systémů pro strojový překlad" **Random**: "how to use"→"jak ho lze využít", "train"→"trénování", "simple"→"jednoduchých"
Chinese→ English	**Source**: "凌寒再次挥手，又结结实实地抽了他一巴掌。" **Reference**: Ling Han raised his hand once more, and gave him another solid slap. **Proper**: '凌寒"→"Ling Han" **Random**: '手"→"his hand"

Table 2: Examples from the WMT 2023 Terminology Shared Task test dataset, based on MuchMore Springer Bilingual Corpus, Rosa and Zouhar (2022), and Jiang et al. (2023). **Base** is without any terminologies, **Proper** is real terminologies and **Random** are random but aligned phrases from source to the reference.

2 Task Description

We focus on how translation quality improves with the incorporation of segment-level terminology on German→English, English→Czech, and Chinese→English datasets. Participants are given source sentences along with a segment-level terminology dictionary (*Source* and *Hints* in Table 1). For the purposes of this study, we define terminology as low-frequency words or phrases that occur typically within a particular domain, such as computer science paper abstract. We scan the source and references for such phrases and provide this segment-level annotation, together with the source, to the participants in the form $X \rightarrow Y$ where X is a span from the source and Y is a span from the reference (*Proper* in Table 2).

Given that the participants are given a part of the reference, Y, this raises the following question: *Is the improvement in translation quality due to the information that a particular terminology X is translated as Y or merely because a part of the reference is leaked to the model?* To better attribute any performance gains, we therefore also test a different mode, where we give the participants "terminologies" where X' and Y' are still aligned spans and translations of each other, but sampled randomly. That is, they are treated as terminology but are, in fact, random phrases (*Random* in Table 2). For this reason, we ask the participants to carry out the translation in three distinct modes:

- **Base**: MT with no terminology dictionary.
- **Proper**: MT with a terminology dictionary. For example *"Sprachmodell" → "language model"*.
- **Random**: MT with randomly chosen, but correct, non-terminological translations. For example *"Hund" → "dog"*.

By comparing performance across these modes, we isolate the model's inherent translation ability and its ability to make use of the terminology.

3 Data

For MT training, the participants were restricted to only the parallel or monolingual datasets enumerated in the WMT general track (Kocmi et al., 2023).[2] The inclusion of pre-trained models was permitted, provided that such usage was explicitly declared. Any employment of terminology-specific datasets that were not part of the specified resources was expressly disallowed. For the terminology-targeted evaluation, we repurposed one dataset and created two new ones. From all of them, we provided 100 segments to the participants as a sanity-check development set. See examples for all language pairs in Table 2.

3.1 Chinese→English Test Data

Our Chinese→English translation test data is sourced from the BWB corpus (Jiang et al., 2023), which covers web novels annotated with, among others, terminologies. The BWB corpus comprises

[2] www2.statmt.org/wmt23/translation-task.html

X→Y	Count	X/Y Words	Terms
German→English	2963	22.2/22.6	3.8
English→Czech	3005	25.6/21.6	3.6
Chinese→English	2640	9.7/36.9	1.1

Table 3: Our test dataset size, average number of words per line and average number of terms per segment (equal between *Proper* and *Random*).

Term.	Prompt
Proper	Identify and annotate all terminology entities (consider only consecutive words) from source sentence and match them with the counterpart in the translated sentence.
Random	Identify and annotate as many as possible aligned words (consider only consecutive words) between source sentence and the translated sentence.

Prompt 1: The upper prompt formulation extracts proper terminology and the bottom extracts random terminology. See Prompt 2 (Appendix) for the full example with few-shot examples.

~3k sentences across six web novels. These annotations identify each named entity and concept in the sentences, highlighting their co-referred expressions. The average terminology count per line is 1.1 (Table 3). Examples of such terminology are in Table 2. Terminology often faces issues of mistranslation or contextually inconsistent translation. Additionally, MT quality declines when terminology is positioned as the subject due to the Chinese's subject-dropping nature.

3.2 English→Czech and German→English Test Data

For the next two language directions we created a new semi-automatically annotated corpus of terminologies. For English→Czech we used 3k sentence pairs from a dataset of NLP papers abstracts (Rosa and Zouhar, 2022). For German→English we used 3k sentence pairs from a dataset of medical paper abstracts (MuchMore Springer Bilingual Corpus). In both cases, the focus on academic texts was guided by the high occurrence of terminology in this domain (3.8 and 3.6, Table 3).

Automatic alignment tools usually have lower precision than linguists and linguists have lower recall and the collection is both time and budget consuming. Therefore, to extract the aligned terminology, we use human-machine collaboration. First, we use GPT-4 (OpenAI, 2023) to create aligned terminology pairs from source and references. We use two few-shot prompts to collect the raw alignments (Prompts 1 and 2). Then we ask linguists to validate these alignments and fix those that are incorrect (either missing terminology, wrong alignment or pairs that are not a terminology). For the Czech-English language pair, humans revised approximately 8% of GPT annotations. There is no modification to terminology in the German-English GPT annotations. Consultation with German linguist affirmed that no adjustments were necessary. Nonetheless, further examination is needed to fully assess GPT's proficiency in terminology alignment for German. This task was sponsored by Microsoft and we release both the pre- and post-alignment data for the further research of GPT capabilities.

4 Participants and System Descriptions

We received a total of 21 per-language submissions from 7 teams. We provide short descriptions of their systems, based on the submitted details.

AdaptTerm (Moslem et al., 2023b). The terminology-enriched MT system builds on Moslem et al. (2023a); Haque et al. (2020). It consists of:

1. using an LLM to generate bilingual synthetic data based on the provided terminology;
2. fine-tuning a generic model, OPUS, with a mix of the terminology-based synthetic data generated by #1 and a randomly sampled portion of the original generic data; and
3. generating translations with the fine-tuned model from #2, and then fixing translations that do not include the required terms with an LLM.

Lingua Custodia (Liu, 2023). This submission includes all three language directions. They use two strategies to extract synthetic terminology from the training data. The first one relies on the invariable n-grams between the source and the target sentence, while the second one extracts parallel sentences that appear inside another training sample as one terminology item. Then, they train a Transformer-based model with annotated data using the extracted terminology, identical to Alam et al. (2021). In addition, after the text annotation, they further apply several annotated data filters to reduce some bias introduced by the automatic annotation. The final trained model can be used directly to translate a text with any new terminology.

OPUSCAT (Nieminen, 2023). A standard Transformer system is finetuned with parallel data where parts of the source sentences have been annotated with their corresponding translations in the target sentences, causing the system to learn to copy the annotated target parts from the source sentence into the target sentence. The translations are generated using a series of models, with different fine-tuned terminology models acting as backoff models to the base transformer model, in cases where the base transformer output does not contain the specified terminology.

UEDIN (Bogoychev and Chen, 2023). Their primary system, *twoshot*, is 2-shot decoding where we enforce terminology constraints via terminology hints in the source and if this does not work we use alignment-based methods to identify the mistranslated terminology word on the target side and penalize it, giving the decoder a chance to generate the hinted word. System *Tag* is decoding with terminology hints while *LLM* is an unconstrained contrastive system.

BJTU-LB (no description paper). They train the in-context learning ability of the model, and then concatenate the term translation pairs in front of the sentence to be translated as the context. The model can generate different translation results according to different contexts.

VARCO-MT (Park et al., 2023). The *ForceGen* is a Transformer-based model that is tailored to ensure the appearance of given terminology in the generated output. By modifying the input format and decoding process, it incorporates a copy mechanism on the source side, allowing it to copy the target terminology from the provided terminology pairs. During the generation process, it uses a force decoding technique, which compels the model to actively generate the target terminology as needed. The *TSSNMT* is a novel Transformer-based NMT model that uses a shared encoder to process both input text and terminology. The model then employs cross-attention mechanisms between the two encoder hidden states and passes them through a gate, enabling the model to autonomously decide which pieces of information (input or terminology) to focus on during translation.

Huawei. Did not submit system description. The translations are also on a subset not used for final evaluation. We include the results in the analysis sections in gray (Huawei) for completeness but urge the readers not to draw any comparisons to other systems.

5 Evaluation

Our evaluation is focused on: (1) general translation quality, (2) quality of translation of specific terminologies, and (3) efficiency in using segment-level terminology dictionaries.

Standard Metrics. Following recent trends in MT evaluation (Kocmi et al., 2021), we use ChrF (Popović, 2015) and COMET (Rei et al., 2020) for the general translation quality evaluation.[3] While the latter one is generally touted as more robust and correlated more with human judgement, in this case we are also concerned in exact match of n-grams, which is captured by ChrF.

Term Success Rate. In the terminology success rate we compare the machine-translated terms with their dictionary equivalents. One would be tempted to check for the presence of the reference terminology translation in the output by the regular expression match. However, this is sensitive to minor orthographic variants. Therefore, we use fuzzy search with threshold of 90% to scan for terminology matches, yielding a number between 0 (no terminology translated correctly) to 1 (all terminology translated correctly).

Term Consistency. This metric looks at whether technical terms are translated uniformly across the entire text corpus. We aim for high consistency, measured by the low occurrence of multiple translations for the same term within the text. We use the approach suggested by Semenov and Bojar (2022). Given the source sentences, outputs, and source terms assigned to each sentence, we firstly make word alignment for the source sentences and outputs, and extract the aligned translated terms for each source term occurrence. Then, we automatically choose the "pseudo-reference" terminology translations, based on which translation of which source term occurred in the text first. In the last step, we compare two sets –the real outputs and the pseudo-references for each term occurrence– by means of F_1 score on a scale of 0 (no consistent terminology) to 1 (all terminology translated consistently).

[3] ChrF uses the defaults from sacreBLEU (Post, 2018) and COMET is `wmt22-comet-da`.

System	ChrF De→En	ChrF En→Cs	ChrF Zh→En
AdaptTerm	61.0	64.4	37.5
Lingua Custodia	61.8⋆	67.7⋆	32.6
OPUS-CAT	53.6	62.5	24.5
UEDIN$_{LLM}$	60.0	64.8	41.2
UEDIN$_{Tag}$	58.3	64.7	41.0
UEDIN$_{Twoshot}$	60.5	62.4	34.5
BJTU-LB			43.8⋆
VARCO-MT$_{TSSNMT}$			43.0
VARCO-MT$_{ForceGen}$			40.5
Huawei	62.1	58.2	36.8

System	COMET$^{DA}_{22}$ De→En	COMET$^{DA}_{22}$ En→Cs	COMET$^{DA}_{22}$ Zh→En
AdaptTerm	0.801	0.841	0.688
Lingua Custodia	0.735	0.834	0.609
OPUS-CAT	0.790	0.869⋆	0.521
UEDIN$_{LLM}$	0.813⋆	0.869⋆	0.757⋆
UEDIN$_{Tag}$	0.809	0.868	0.757⋆
UEDIN$_{Twoshot}$	0.792	0.835	0.650
BJTU-LB			0.751
VARCO-MT$_{TSSNMT}$			0.755
VARCO-MT$_{ForceGen}$			0.715
Huawei	0.843	0.887	0.666

Table 4: Averages of ChrF and COMET scores with *Proper* terminology dictionaries. The ⋆ marks best within each column (language) and metric.

5.1 Main Results (Table 4)

We begin the comparison using two standard metrics of MT quality in the case where *Proper* terminology dictionaries were provided. The choice of the best-performing system diverges based on the two metrics: *Lingua Custodia* is selected as the best by ChrF in two language directions, it ranks the same system on Zh→En as the second lowest-performing one. In contrast, COMET ranks *UEDIN$_{LLM}$* as the best across all three language directions. Given that this metric better captures human judgement (Freitag et al., 2022), this ranking is likely more close to the true quality.

5.2 Terminology Quality (Table 6)

The results are even more different when focusing solely on the correctness of the terminology. Overall, most systems translate 60%-70% of terminologies correctly. For terminology consistency, the most immediate outlier is *VARCO-MT$_{TSSNMT}$*, yielding impressive score of 0.971 on Chinese→English. Table 5 illustrates how even in the same document the terminology can be translated differently, which is undesired.

Source	Die Krankheit entwickelt sich bei Kindern und jungen Erwachsenen und folgt dem Muster der Blaschko-Linie.
MT	The condition develops during childhood and adolescence and follows the pattern of the blaschko line.
	...
Source	Ungefähr 95% aller Personen, die M. leprae ausgesetzt sind, entwickeln die Krankheit nicht.
MT	About 95% of all individuals exposed to M. leprae do not develop the disease.

Table 5: Example of term inconsistency (*Krankheit → disease, condition*) within the same document.

System	Terminology Consistency De→En	Terminology Consistency En→Cs	Terminology Consistency Zh→En
AdaptTerm	0.617	0.753	0.750
Lingua Custodia	0.602	0.766	0.696
OPUS-CAT	0.661⋆	0.808⋆	0.293
UEDIN$_{LLM}$	0.588	0.741	0.713
UEDIN$_{Tag}$	0.606	0.750	0.755
UEDIN$_{Twoshot}$	0.574	0.737	0.622
BJTU-LB			0.764
VARCO-MT$_{TSSNMT}$			0.971⋆
VARCO-MT$_{ForceGen}$			0.773
Huawei	0.788	0.603	0.562

System	Terminology Success Rate De→En	Terminology Success Rate En→Cs	Terminology Success Rate Zh→En
AdaptTerm	0.587	0.613	0.758
Lingua Custodia	0.622⋆	0.662	0.747
OPUS-CAT	0.443	0.557	0.124
UEDIN$_{LLM}$	0.560	0.629⋆	0.753
UEDIN$_{Tag}$	0.539	0.626	0.739
UEDIN$_{Twoshot}$	0.587	0.562	0.536
VARCO-MT$_{TSSNMT}$			0.779
VARCO-MT$_{ForceGen}$			0.800⋆
BJTU-LB			0.749
Huawei	0.694	0.462	0.486

Table 6: Averages of Terminology Consistency and Terminology Success Rate with *Proper* terminology dictionaries. The ⋆ marks best within each column (language) and metric.

5.3 Terminology Utility (Tables 7 and 8)

Previous investigations into the general translation and terminology translation quality did not reveal many differences between the systems. We now focus on the usefulness of the additional information and show the difference between *Base* and either *Proper* or *Random* terminology dictionaries in Table 7. Notably *AdaptTerm* and *Lingua Custodia* improve the most from their *Base* version. With an exception of *OPUS-CAT*, both ChrF and COMET

System	ChrF +Proper	ChrF +Random	$COMET_{22}^{DA}$ +Proper	$COMET_{22}^{DA}$ +Random	T. Consistency +Proper	T. Consistency +Random	T. Success Rate +Proper	T. Success Rate +Random
AdaptTerm	9.0	11.6	0.043	0.054	0.020	-0.120	0.239	0.257
Lingua Custodia	10.1	11.8	0.032	0.026	0.118	-0.059	0.345	0.341
OPUSCAT	-10.2	-1.0	-0.031	0.012	0.055	-0.044	-0.285	0.074
$UEDIN_{LLM}$	6.4	7.5	0.011	0.017	0.027	-0.100	0.164	0.128
$UEDIN_{Tag}$	5.4	6.5	0.010	0.013	0.055	-0.090	0.162	0.117
$UEDIN_{Twoshot}$	6.9	5.9	0.029	0.012	0.045	-0.074	0.193	0.297
BJTU-LB †	2.5	0.8	0.015	0.007	0.058	-0.150	0.178	-0.015
$VARCO\text{-}MT_{TSSNMT}$ †	8.3	4.7	0.054	0.017	0.171	-0.189	0.515	0.508
$VARCO\text{-}MT_{ForceGen}$ †	3.4	0.9	0.019	0.003	0.166	-0.137	0.417	0.202
Huawei	0.2	0.9	-0.004	0.010	-0.010	0.042	0.033	-0.012

Table 7: Average difference in each metric between the *Base* and added dictionary (*Proper* or *Random*). All numbers are averages across all languages except for † which is Chinese→English only.

improves across all metrics when given any of the two dictionaries. This challenges the notion that the additional information supplied to the MT system needs to be terminology while in fact it can be any information that leaks from the reference. Focusing on a particular language pair in Table 8, there seems to be weak effect of lower variance when terminology dictionaries are provided.

System	$COMET_{22}^{DA}$ Zh→En Base	Proper	Random
AdaptTerm	$0.638_{0.142}$	$0.688_{0.109}$	$0.678_{0.104}$
Lingua Custodia	$0.476_{0.148}$	$0.609_{0.128}$	$0.528_{0.124}$
OPUSCAT	$0.557_{0.147}$	$0.521_{0.155}$	$0.624_{0.132}$
$UEDIN_{LLM}$	$0.750_{0.076}$	$0.757_{0.075}$	$0.753_{0.078}$
$UEDIN_{Tag}$	$0.747_{0.083}$	$0.757_{0.077}$	$0.747_{0.083}$
$UEDIN_{Twoshot}$	$0.572_{0.158}$	$0.650_{0.121}$	$0.596_{0.155}$
BJTU-LB	$0.736_{0.101}$	$0.751_{0.092}$	$0.743_{0.092}$
$VARCO\text{-}MT_{TSSNMT}$	$0.701_{0.145}$	$0.755_{0.138}$	$0.718_{0.135}$
$VARCO\text{-}MT_{ForceGen}$	$0.696_{0.094}$	$0.715_{0.091}$	$0.699_{0.095}$
Huawei	$0.679_{0.101}$	$0.666_{0.104}$	$0.709_{0.103}$

Table 8: Distribution of segment-level COMET scores on Chinese→English language direction (if available) between all three translation modes. Notation: mean $_{var}$.

6 Related Work

Similar to the previously shared task on translation using terminologies (Alam et al., 2021), our terminology hints are mined semi-automatically. We also extend this line of work by contrasting random and proper terminologies. The focus on terminologies in translation is an important one. Both Zouhar et al. (2020) and Semenov and Bojar (2022) show that the ordering of the system diverges when comparing performance on terminologies versus general performance.

Constrained Decoding. A simple paradigm for improving terminology translation is constrained decoding. Anderson et al. (2017) track constraint satisfaction using a finite-state machine. Hokamp and Liu (2017) reduce the time complexity to linear and Post and Vilar (2018) further improve on this.

Other approaches. Other than constrained decoding, several works have approached the problem by guiding the text generation model, including those that modify the token-level distribution using an external model (Stahlberg et al., 2017; Gulcehre et al., 2017; Chatterjee et al., 2017; Pascual et al., 2021), and those that incorporate constraints into the training process through additional annotations (Dinu et al., 2019; Bergmanis and Pinnis, 2021; Niehues, 2021, *inter alia*).

7 Conclusion

This iteration of machine translation with terminologies focused on evaluating the efficiency of using segment-level terminology dictionaries. I.e. it is not enough that the system performs well but it should also perform better when given this additional information. Indeed, the improvement between *Base* and *Proper* terminology enriched translations ranged across systems between 0 and 10 ChrF points. This helps isolate which terminology-enhancement methods are the most useful.

Limitations

The evaluation datasets are based on publicly-available data, which might have been leaked to the training of submitted systems, skewing the results. We further acknowledge that the comparisons in this work were not done using statistical testing.

Ethical Consideration

The work of both linguist working on the validation of GPT alignment was well-paid of around a twice to three times the minimal hourly wage in their respective countries. The annotated texts did not contain any sensitive or explicit passages.

References

Md Mahfuz Ibn Alam, Ivana Kvapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Kweonwoo Jung, Philipp Koehn, and Vassilina Nikoulina. 2021. Findings of the WMT shared task on machine translation using terminologies. In *Proceedings of the Sixth Conference on Machine Translation*, pages 652–663, Online. Association for Computational Linguistics.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2017. Guided open vocabulary image captioning with constrained beam search. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 936–945, Copenhagen, Denmark. Association for Computational Linguistics.

Mihael Arcan, Marco Turchi, Sara Tonelli, and Paul Buitelaar. 2017. Leveraging bilingual terminology to improve machine translation in a CAT environment. *Natural Language Engineering*, 23(5):763–788.

Toms Bergmanis and Mārcis Pinnis. 2021. Facilitating terminology translation with target lemma annotations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3105–3111, Online. Association for Computational Linguistics.

Nikolay Bogoychev and Pinzhen Chen. 2023. Terminology-aware translation with post-processing using constrained decoding and large language models. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

M Teresa Cabré. 2010. Terminology and translation. *Handbook of translation studies*, 1:356–365.

Rajen Chatterjee, Matteo Negri, Marco Turchi, Marcello Federico, Lucia Specia, and Frédéric Blain. 2017. Guiding neural machine translation decoding with external knowledge. In *Proceedings of the Second Conference on Machine Translation*, pages 157–168, Copenhagen, Denmark. Association for Computational Linguistics.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, and Yoshua Bengio. 2017. On integrating a language model into neural machine translation. *Computer Speech & Language*, 45:137–148.

Rejwanul Haque, Yasmin Moslem, and Andy Way. 2020. Terminology-aware sentence mining for NMT domain adaptation: ADAPT's submission to the adap-MT 2020 English-to-Hindi AI translation shared task. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON): Adap-MT 2020 Shared Task*, pages 17–23, Patna, India. NLP Association of India (NLPAI).

Eva Hasler, Adrià de Gispert, Gonzalo Iglesias, and Bill Byrne. 2018. Neural machine translation decoding with terminology constraints. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 506–512, New Orleans, Louisiana. Association for Computational Linguistics.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

Md Mahfuz ibn Alam, Antonios Anastasopoulos, Laurent Besacier, James Cross, Matthias Gallé, Philipp Koehn, and Vassilina Nikoulina. 2021. On the evaluation of machine translation for terminology consistency.

Yuchen Eleanor Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Mrinmaya Sachan, and Ryan Cotterell. 2023. Discourse-centric evaluation of document-level machine translation with a new densely annotated parallel corpus of novels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7853–7872, Toronto, Canada. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and Maja Popović. 2023. Findings of the 2023 conference on machine

translation (WMT23). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Jingshu Liu. 2023. Lingua custodia's participation at the WMT 2023 terminology shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023a. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, John D. Kelleher, Rejwanul Haque, and Andy Way. 2023b. Domain terminology integration into machine translation: Leveraging large language models. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Jan Niehues. 2021. Continuous learning in neural machine translation using bilingual dictionaries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 830–840, Online. Association for Computational Linguistics.

Tommi Nieminen. 2023. Opus-cat terminology systems for the wmt23 terminology shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report.

Geon Woo Park, Junghwa Lee, Meiying Ren, Allison Shindell, and Yeonsoo Lee. 2023. VARCO-MT: NCSOFT's WMT'23 terminology shared task submission. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Damian Pascual, Beni Egressy, Clara Meister, Ryan Cotterell, and Roger Wattenhofer. 2021. A plug-and-play method for controlled text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3973–3997, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Rudolf Rosa and Vilém Zouhar. 2022. Czech and English abstracts of ÚFAL papers (2022-11-11). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Kirill Semenov and Ondřej Bojar. 2022. Automated evaluation metric for terminology consistency in MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 450–457, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Felix Stahlberg, Adrià de Gispert, Eva Hasler, and Bill Byrne. 2017. Neural machine translation by minimising the Bayes-risk with respect to syntactic translation lattices. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 362–368, Valencia, Spain. Association for Computational Linguistics.

Chelo Vargas-Sierra. 2011. Translation-oriented terminology management and ICTs: present and future. *Interdisciplinarity and languages: Current Issues in Research, Teaching, Professional Applications and ICT.*, pages 45–64.

Vilém Zouhar, Tereza Vojtěchová, and Ondřej Bojar. 2020. WMT20 document-level markable error exploration. In *Proceedings of the Fifth Conference on Machine Translation*, pages 371–380, Online. Association for Computational Linguistics.

Identify and annotate all terminology entities (consider only consecutive words) from source sentence and match them with the counterpart in the translated sentence.

_____
Source 'en': after blowing your nose, coughing or sneezing.
Translation 'fr': après s'être mouché ou avoir toussé/éternué.
Annotation: {'en': 'coughing', 'fr': 'toussé'}, {'en': 'sneezing', 'fr': 'éternué'}

_____
Source 'zh': 仙羽郡，武宗学府，后山林中，一个身披宽松武袍的削瘦少年，双盘下蹲，舌尖抵住牙齿，全身力量集中于左右两拳，轰打人粗大树。
Translation 'en': In mountainous forest behind Xianyu prefecture , martial arts training institute , there was thin young man wearing loose and comfortable martial artist robe . In the lotus position with his tongue against his teeth, he focused all his strength into both his fists and pummeled huge tree.
Annotation: {'zh':' 仙羽郡','en': 'Xianyu prefecture'}, {'zh':' 武宗学府','en': 'a martial arts training institute'}

_____
Source 'en': According to Statistics Austria's current estimate from April 2015, expenditure for research and development carried out in Austria in 2015 is projected to grow nominally by around €271.36 million or 2.76% compared to 2014, thereby exceeding the €10 billion threshold for the first time (€10.10 billion).
Translation 'de': Gemäß der aktuellen Globalschätzung der Statistik Austria vom April 2015 werden die gesamten Ausgaben für Forschung und Entwicklung in Österreich 2015 voraussichtlich gegenüber dem Jahr 2014 um rd.271,36 Mio. €bzw. 2,76% nominell wachsen und damit erstmals die 10 Mrd. €-Schwelle überschreiten (10,10 Mrd. €).
Annotation: {'en': 'expenditure', 'de': 'gesamten Ausgaben'}, {'en': 'research and development', 'de': 'Forschung und Entwicklung'}, {'en': 'threshold', 'de': '-Schwelle'}

_____
Source 'cs': Podle ředitele Institutu veřejné správy Filipa Hrůzy si pořadatelé nyní musí vyhodnotit, jestli je pro Brno závod výhodný.
Translation 'en': According to the Head of the Public Administration Institute, Filip Hrůza, the organizers must now assess whether the race brings benefits to Brno.
Annotation: {'en': 'Public Administration Institute', 'cs': 'Institutu veřejné správy'}, {'en': 'race', 'cs': 'závod'}

_____
Source '**{source_lang}**': **{source_segment}**
Translation '**{target_lang}**': **{translated_segment}**
Annotation:

Prompt 2: The prompt for collecting aligned terminology with GPT-4. **Bolded** text is replaced with current segment.

Findings of the WMT 2023 Shared Task on Automatic Post-Editing

Pushpak Bhattacharyya
IIT Bombay
pb@cse.iitb.ac.in

Rajen Chatterjee
Apple Inc.
rajen_c@apple.com

Markus Freitag
Google
freitag@google.com

Diptesh Kanojia
University of Surrey
d.kanojia@surrey.ac.uk

Matteo Negri
Fondazione Bruno Kessler
negri@fbk.eu

Marco Turchi
Zoom Video Communications
marco.turchi@zoom.us

Abstract

We present the results from the 9^{th} round of the WMT shared task on MT Automatic Post-Editing, which consists of automatically correcting the output of a "black-box" machine translation system by learning from human corrections. Like last year, the task focused on English→Marathi, with data coming from multiple domains (healthcare, tourism, and general/news). Despite the consistent task framework, this year's data proved to be extremely challenging. As a matter of fact, none of the official submissions from the participating teams succeeded in improving the quality of the already high-level initial translations (with baseline TER and BLEU scores of 26.6 and 70.66, respectively). Only one run, accepted as a "late" submission, achieved automatic evaluation scores that exceeded the baseline.

1 Introduction

This paper presents the results of the 9^{th} round of the WMT task on MT Automatic Post-Editing (APE). The task involves the automatic correction of the output generated by a "black-box" machine translation system by learning from human-revised machine-translated output supplied as training material. The overall task formulation (see Section 2) remained consistent with that of all previous rounds. In this formulation, the challenge revolves around fixing errors in English documents that have been automatically translated by a state-of-the-art, non-domain-adapted neural MT (NMT) system unknown to the participants. In continuity with last year's round, the evaluation focused on English→Marathi,[1] with training/dev/test data selected from a mix of domains, namely- healthcare, tourism, and general/news (see Section 3).

Three teams participated in the task by submitting a total of four runs for the final evaluation (see Section 4).[2] However, while only two out of the three participants were able to submit their runs on time, the one remaining submission arrived with a two-month delay. This led us to categorize it as a late (therefore, unofficial) submission for the sake of fairness to the other participants.

For all the teams, the task posed significant challenges primarily due to the high average quality of the initial translations slated for post-editing (26.6 TER / 70.66 BLEU / 79.78 chrF). This challenge was compounded by the substantial imbalance in distribution between near-perfect translations (approximately 40% of the total) and those necessitating extensive revisions (approximately 20%). As a consequence, none of the official runs was able to improve over the baseline in terms of the task's automatic evaluation metrics (Section 5.1), with the best run achieving results (27.73 TER / 69.03 BLEU / 78.64 chrF) that highlight a slight quality degradation compared to the original, untouched NMT outputs that represent our baseline. For the sake of completeness, we report that the late submission achieved a slight improvement over the baseline, attested by TER, BLEU, and chrF scores of 25.74, 71.27, and 80.41, respectively. The results computed by means of automatic evaluation metrics were also confirmed by our human evaluation based on direct assessment (Section 5.2).

2 Task Description

MT Automatic Post-Editing (APE) is the task of automatically correcting errors in a machine-translated text. As pointed out by (Chatterjee et al., 2015), from the application point of view, the task is motivated by its possible uses to:

- Enhance MT output by harnessing information that is not available to the decoder or by conducting deeper text analysis, which may

[1] Marathi is an Indo-Aryan language predominantly spoken by Marathi people in the Indian state of Maharashtra.

[2] A fourth participant withdrew the submitted run, which was affected by major errors in the generated outputs.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 672–681
December 6–7, 2023.

be prohibitively expensive during the decoding phase.

- Address systematic errors stemming from an MT system whose decoding process is inaccessible for focused modifications.
- Provide professional translators with improved MT output quality, thereby reducing the need for subsequent human post-editing.
- Tailor the output of a general-purpose MT system to align with the lexicon and style requirements of a specific application domain.

This 9^{th} round of the WMT APE shared task kept the same overall evaluation setting of the previous eight rounds. Specifically, the participating systems had to automatically correct the output of an unknown "black-box" MT system (a generic NMT system not adapted to the target domain) by learning from training data containing human revisions of translations produced by the same system. For the second year in a row, the selected language pair was English-Marathi (with Marathi as the target language for post-editing). Training, development and test data were drawn from the following three domains: healthcare, tourism, and general/news.

3 Data, Metrics, Baseline

3.1 Data

In continuity with last year, the selected language pair is English-Marathi. Marathi is one of the most spoken Indian languages, with approximately 83 million native speakers and 16 million speakers as a second/third language.[3] Marathi is a known agglutinative language and presents various challenges to machine translation compared to its other Indian counterparts (Khatri et al., 2021; Banerjee et al., 2021). Moreover, the English-Marathi language pair is considered low-resource compared to English-Hindi/Bengali/Malayalam (Ramesh et al., 2022), despite having more native speakers worldwide.

The **training** and **development** datasets supplied to the participants remain consistent with those employed in the 2022 iteration of the task. These datasets consist of 18,000 and 1,000 (*source*, *target*, *human post-edit*) triplets, wherein:

- The source (SRC) is an English sentence;
- The target (TGT) is a Marathi translation of the source produced by a generic, black-box NMT system unknown to participants. This multilingual NMT system (Ramesh et al., 2022) is based on the Transformer architecture (Vaswani et al., 2017) and is trained on a total of 49 million sentence pairs where the En-Mr parallel corpus is 4.5 million sentence pairs. This parallel data is generic and covers many domains, including the three domains covered by the APE 2023 test set, namely-healthcare, tourism/culture and general/news.
- The human post-edit (PE) is a manually revised version of the target, which was produced by native Marathi speakers.

We provide the same corpus of artificially generated data as additional training material from the last round. It consists of 2 million triplets derived from the *Anuvaad* en-mr parallel corpus.[4] The *Anuvaad* parallel corpus consists of data for 12 en-X language pairs, where X comprises 12 Indian languages, including Marathi. The English-Marathi data consists of 2.5 million parallel sentences. Specifically, the *source*, *target*, *post-edit* instances of this synthetic corpus are respectively obtained by combining: *i)* the original English source sentence from the *Anuvaad* corpus, *ii)* its automatic translation into Marathi,[5] *iii)* the original Marathi target sentence from the *Anuvaad* corpus.

Test data consisted of 1,000 (*source*, *target*) pairs, similar in nature to the corresponding elements in the train/dev sets (*i.e.,* same domains, same NMT system). The human post-edits of the target elements were left apart to measure APE systems' performance both with automatic metrics (TER, BLEU, chrF) and via human evaluation.

3.2 Metrics

The participating systems were evaluated both by means of automatic metrics and manually. Automatic evaluation (Section 5.1) was carried out after tokenizing the data with *sacremoses*[6], by computing the distance between the automatic post-edits produced by each system for the target elements of

[3] Ethnologue-2022 - Ethnologue has been an active research project since 1951 which maintains online archives of recognized languages list, and their statistics.

[4] https://github.com/project-anuvaad/anuvaad-parallel-corpus

[5] from IndicTrans En-X Model (Ramesh et al., 2022)

[6] https://pypi.org/project/sacremoses/

the test set, and the human corrections of the same test items.

The official systems' ranking is based on the average (case-sensitive) TER (Snover et al., 2006) calculated on the test set by using the TERcom[7] software: lower average TER scores correspond to higher ranks. As additional performance indicators, BLEU (Papineni et al., 2002) and chrF (Popović, 2015) were computed[8]. The human evaluation (Section 5.2) was conducted via source-based direct human assessment (Graham et al., 2013a).

3.3 Baseline

The official baseline results were the TER/BLEU/chrF scores calculated by comparing the raw MT output with human post-edits. This corresponds to the score achieved by a "*do-nothing*" APE system that leaves all the test targets unmodified.

4 Submissions

As shown in Table 1, this year, we received submissions from three teams, one of which submitted their run with a two-month delay that motivates its categorization as a late submission[9]. The main characteristics of the participating systems are summarized below.

Korea Advanced Institute of Science and Technology (kaistai). This team participated with a system inspired by the recent surge of large language models (LLMs) that have been successfully applied to a variety of language generation tasks. Their goal was to verify whether LLMs could perform the APE task through prompting. To this aim, they used gpt-3.5-turbo with specific prompts designed to generate either (a) post-edits or (b) post-edits along with the rationales behind them. While the results of preliminary evaluations based on COMET suggested the viability of the approach for medium-/high-resource language pairs, they also highlighted that the often radical changes produced by LLMs can potentially be penalized by more strict reference-based evaluations based on BLEU, TER, or chrF.

Korea University (KU_UPs). The participation of this team was centred on data filtering techniques. With a focus on removing potentially harmful material from a model training perspective, the proposed method concentrates on eliminating the two extremes of the training data distribution: the (near-)perfect MT outputs on one side, and those that require complete rewriting on the other. According to preliminary experiments carried out on previous APE datasets (WMT2020/2021/2022), data selection driven by TER and COMET yields better performance when the outlier instances requiring excessive post-editing are removed from the training. On this basis, the submitted APE system was built by training the multilingual M2M100-418M model (Fan et al., 2021).

Huawei Translation Service Center and Xiamen University School of Informatics (HW_TSC). – late submission – This team participated with a Transformer-based system pre-trained on the provided synthetic APE data and then fine-tuned on the real APE data augmented via automatic translation (with Google Translate run on the post-edits in the training set) and by integrating En-Mr parallel sentences from FLORES-200 (NLLB Team et al., 2022). R-Drop (Liang et al., 2021), which regularizes the training inconsistency induced by dropout, is used to mitigate overfitting during the training phase. A sentence-level Quality Estimation system is also used to select the most appropriate output, choosing between the original translation and the corresponding APE-generated output.

5 Results

5.1 Automatic Evaluation

Automatic evaluation results are shown in Table 2. The submitted runs are ranked based on the average TER (case-sensitive) computed using human post-edits of the MT segments as a reference, which is the APE task's primary evaluation metric. To provide a broader view of systems' performance, BLEU and chrF results computed using the same references are also reported.

As can be seen from the table, the three rankings coherently show that the best official submission (by the KU_UPS team, which achieved scores of 27.73 TER, 69.03 BLEU, and 78.64 chrF) outperforms the others. None of them, however, was able to improve the quality of the original translations (*i.e.* the *do nothing* baseline), differently from

[7] http://www.cs.umd.edu/~snover/tercom/

[8] chrF was computed using SacreBLEU https://pypi.org/project/sacrebleu/(version 2.3.0)

[9] A fourth participating team retracted their submitted run due to errors in the generated outputs that significantly affected their final results.

ID	Participating team
kaistai	Korea Advanced Institute of Science and Technology, South Korea
KU_UPs	Korea University, South Korea (Moon et al., 2023)
HW_TSC	Huawei Translation Service Center & Xiamen University School of Informatics, China (Yu et al., 2023)

Table 1: Participants in the WMT23 Automatic Post-Editing task.

		TER	BLEU	CHRF
en-mr	HW-TSC_HW_1_PRIMARY.txt†	25.74	71.27	80.41
	baseline (MT)	26.60	70.66	79.78
	KU_UPs-filtered4-PRIMARY.tsv	27.73	69.03	78.64
	kaistai_prompt-wo-cot_contrastive	54.59	40.97	67.24
	kaistai_prompt-w-cot_primary	58.55	31.63	61.61

Table 2: NEWResults for the WMT23 APE English-Marathi shared task – average TER (↓), BLEU (↑), chrF (↑). Gray† indicates a late submission, which was received after the conclusion of this year's human evaluation and, consequently, is not discussed in Section 5.2.

the slightly better outputs of the late submission by HW_TSC. This prompts further analyses to explore the underlying reasons for this unexpected outcome. We do this in two ways: 1) by giving a closer look at systems' behaviour (Section 5.1.1), in order to spot trends in their post-editing strategies; 2) by analysing the task's inherent level of difficulty (Section 5.1.2) in terms of the possibility to learn valuable correction patterns from the training data and effectively apply them to the supplied test set.

5.1.1 Analysis: Systems' Behaviour

Modified, improved and deteriorated sentences To gain a first insight into the behaviour of participating systems, Table 3 provides an overview of each submitted run, detailing the number of modified, improved, and deteriorated sentences, along with the systems' overall precision (*i.e.,* the ratio of improved sentences to the total count of modified instances where improvement or deterioration is observed). It's worth noting that each system has modified a much higher number of sentences than the combined total of improved and deteriorated ones. This discrepancy accounts for modified sentences in which the corrections do not result in any variations in TER. This "grey area", where the automatic assessment of quality improvement or degradation is not feasible, underscores the importance of including human evaluation for a comprehensive assessment of systems' performance (see Section 5.2). As can be seen from the table, and in line with the findings from previous rounds, conservative post-editing seems to yield better results compared to the adoption of aggressive strategies. The difference between the top-ranked system and the other submitted runs is indeed evident when we look at the proportion of modified test sentences (37.4%[10] vs $\geq 93.1\%$), indicating that limiting the applied edits to the strictly necessary ones remains the main challenge to achieve significant quality improvements. While this outcome may be influenced by the reference-based automatic evaluation framework employed (as it penalizes correct edit operations that deviate from those presented in the reference), it is noteworthy that the results of the manual evaluation, as detailed in Section 5.2, align with this observation.

Another observation is that precision is certainly the other key factor in achieving good APE results. Besides being much more conservative, the best submission stems, in fact, for a higher precision in selecting the edit operations to be applied (48.11[11] vs ≤ 21.00). Also, this finding aligns with the outcomes of previous rounds, in which the winning system consistently exhibited the highest precision. Notably, the precision of this year's official submissions (averaging 29.62) is significantly lower than the values observed in previous rounds (*e.g.,* 69.0, 53.96, 69.49 for the top-ranked system in 2020, 2021, and 2022, respectively).[12] This difference in precision may well explain why none of them were able to improve upon the baseline results.

Edit operations Further indications about the system's behaviour can be drawn from a more fine-

[10] Which drops to 24.4% for the late submission.

[11] Further increased to 51.89 for the late submission.

[12] This holds even if we include the late submission in the computation, with an average precision that slightly grows to 35.19.

Systems	Modified	Improved	Deteriorated	Prec.
HW-TSC_HW_1_PRIMARY.txt†	244 (24.4%)	110 (45.08%)	102 (41.8%)	51.89
KU_UPs-filtered4-PRIMARY.tsv	374 (37.4%)	153 (40.9%)	165 (44.11%)	48.11
kaistai_prompt-wo-cot_contrastive	931 (93.1%)	187 (20.08%)	709 (76.15%)	20.87
kaistai_prompt-w-cot_primary	989 (98.9%)	193 (19.51%)	777 (78.56%)	19.89
Average	76.46%	26.83%	66.27%	29.62

Table 3: Number (raw and proportion) of test sentences modified, improved and deteriorated by each run submitted to the APE 2023 English-Marathi subtask. The "Prec." column shows systems' precision as the ratio between the number of improved sentences and the total number of modified instances for which improvement or deterioration can be assessed in terms of TER variations. Average values considering only the three official submissions.

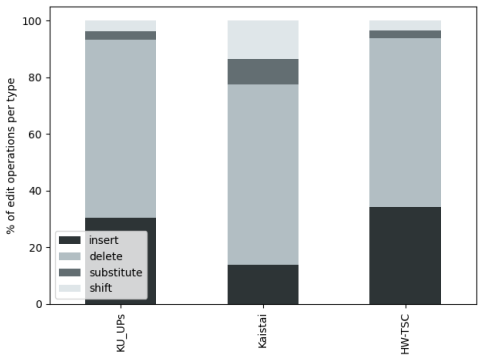


Figure 1: Distribution of edit operations (insertions, deletions, substitutions and shifts) performed by the three primary submissions to the WMT23 APE English-Marathi shared task.

grained analysis of the distribution of their edit operations (insertions, deletions, substitutions, and shifts). Such distribution is obtained by computing the TER between the original MT output and the output of each primary submission, taken as a reference. As shown in Figure 1, although the overall behaviour of the systems is similar, some differences are noticeable. Indeed, in line with previous rounds, they all exhibit a high percentage of deletions, followed by insertions, substitutions and shifts. However, for the best official submission,[13] the percentage of the latter two types of operations is minimal (2.9% substitutions and 3.67% shifts) and balanced by a less skewed distribution of insertions (30.52%) and deletions (62.91%). Especially the comparatively higher proportion of more "radical" (*i.e.,* structural) modifications applied by the worse system (13.43% shifts), which again suggests its lower conservativeness, can account for its lower automatic evaluation scores.

[13]Note, however, that the same consideration also applies for the late submission.

5.1.2 Analysis: Complexity Indicators

While systems' behaviour is influenced by implementation and architectural choices on the one hand, it also depends on the data used for training, development, and evaluation on the other. Therefore, looking at the intrinsic difficulty of the task from a data perspective is also crucial for interpreting the observed performance of the systems. To delve into this aspect, we concentrate on the possibility of learning useful correction patterns during training and successfully applying them at test time. We analyse such a possibility in terms of three indicators, namely: *i)* repetition rate, *ii)* MT quality, and *iii)* TER distribution in the test set. For the sake of comparison across the nine rounds of the APE task (2015–2023), Table 4 reports, for each dataset, information about the first two aspects. The third one, instead, will be discussed by referring to Figure 2.

Repetition Rate The repetition rate (RR), measures the repetitiveness inside a text by looking at the rate of non-singleton n-gram types (n=1...4) and combining them using the geometric mean. Higher RR values indicate greater text repetitiveness, which may imply an increased likelihood of learning correction patterns from the training set that are also applicable to the test set. As shown in the last row of Table 4, the RR values for the SRC, TGT, and PE elements (averaged across the training, development, and test sets) are relatively low. Furthermore, upon closer examination, Table 5 reveals a non-negligible difference between the RR values of the SRC, TGT, and PE elements in the training set compared to the corresponding values calculated on the test set. This difference is particularly pronounced for the PE sentences, where the RR is more than two times higher. Although the reported RR values can be considered indicative of a challenging task, it is important to

	Lang.	Domain	MT type	RR_src	RR_tgt	RR_pe	Basel. BLEU	Basel. TER	δ TER
2015	en-es	News	PBSMT	2.9	3.31	3.08	n/a	23.84	+0.31
2016	en-de	IT	PBSMT	6.62	8.84	8.24	62.11	24.76	-3.24
2017	en-de	IT	PBSMT	7.22	9.53	8.95	62.49	24.48	-4.88
2017	de-en	Medical	PBSMT	5.22	6.84	6.29	79.54	15.55	-0.26
2018	en-de	IT	PBSMT	7.14	9.47	8.93	62.99	24.24	-6.24
2018	en-de	IT	NMT	7.11	9.44	8.94	74.73	16.84	-0.38
2019	en-de	IT	NMT	7.11	9.44	8.94	74.73	16.84	-0.78
2019	en-ru	IT	NMT	18.25	14.78	13.24	76.20	16.16	+0.43
2020	en-de	Wiki	NMT	0.65	0.82	0.66	50.21	31.56	-11.35
2020	en-zh	Wiki	NMT	0.81	1.27	1.2	23.12	59.49	-12.13
2021	en-de	Wiki	NMT	0.73	0.78	0.76	71.07	18.05	-0.77
2022	en-mr	health/tourism/news	NMT	1.46	0.89	0.72	67.55	20.28	-3.49
2023	en-mr	health/tourism/news	NMT	1.85	1.24	1.12	70.66	26.60	+1.13

Table 4: Basic information about the APE shared task data released since 2015- languages, domain, type of MT technology, repetition rate and initial translation quality (TER/BLEU of TGT). The last column (δ TER) indicates, for each evaluation round, the difference in TER between the baseline (*i.e.,* the "*do-nothing*" system) and the top-ranked official submission.

Data	RR
Train_src	1.55
Train_mt	1.03
Train_pe	0.81
Dev_src	1.4
Dev_mt	0.8
Dev_pe	0.64
Test_src	2.6
Test_mt	1.9
Test_pe	1.91

Table 5: Repetition Rate (RR) values of source (src), target translation (mt) and post-edited translation (pe) elements in the APE 2023 training, development and test sets.

note that the top-ranked submissions in previous rounds (e.g. in 2022 and 2020) were able to achieve significant improvements over the baseline despite similar RR values (with δ TER values of -3.49 and -12.13, respectively). This variability reinforces the findings from previous rounds, emphasizing that RR alone is insufficient as a complexity indicator. Rather, it underscores the significance of examining its interaction with other indicators and its potential cumulative impact on them.

MT Quality As emphasized by the findings from all previous rounds of the task, a more reliable indicator of complexity is the quality of the machine-translated (TGT) texts that require correction. We assess this quality by computing TER (↓) and BLEU (↑) scores (shown in the Basel. TER/BLEU columns in Table 4), using the human post-edits as references.[14] In principle, higher-quality original translations leave less room for improvement to APE systems, which have at the same time fewer errors to learn from during training and fewer corrections to make at test time. On one side, indeed, training on good (or near-perfect) automatic translations can significantly reduce the number of learned correction patterns. On the other side, testing on similarly high-quality translations can have two effects: *i)* it reduces the number of corrections required and the applicability of learned patterns, and *ii)* it increases the risk of introducing errors, especially when post-editing near-perfect TGTs. This observation is supported by the strong correlation (>0.83) between the initial MT quality ("Basel. TER" in Table 4) and the TER difference between the baseline and the top-ranked submission ("δ TER" in Table 4) previously reported in the analysis of the 2015-2022 rounds by Bhattacharyya et al. (2022).

Looking at the baseline TER score, this year's test data look for a comparatively lower difficulty for APE systems compared to most of the previous rounds, which in only 2 cases (*i.e.,* for the two languages covered in 2020) appear to be less challenging. Interestingly, however, when looking at the baseline BLEU score, the difficulty appears to be higher, with up to 6 previous test sets featuring translations of lower quality (hence easier to handle) compared to this year. The reasons for such differences deserve further investigation, which might shed light on the fact that, contrary to expectations, MT quality is less indicative of this year's task difficulty compared to previous rounds[15].

[14] Scores for the newly introduced chrF metric are not included in the table, as they would not be comparable with values from previous rounds where chrF was not considered.

[15] Considering this year's data, in fact, the correlation between "Basel. TER" and "δ TER" in Table 4 drops from >0.83

TER distribution in the test set Complementary to repetition rate and MT quality, the TER distribution (computed against human references) for the translations present in the test provides valuable insights for interpreting the results of this year's round of the task. While TER distribution and MT quality may appear to be closely related, it's important to note that, even at similar overall quality levels, more or less skewed distributions can create distinct testing conditions. Indeed, as shown by previous analyses (Bojar et al., 2017; Chatterjee et al., 2018, 2019, 2020; Akhbardeh et al., 2021; Bhattacharyya et al., 2022), more challenging rounds of the task were typically characterized by TER distributions heavily skewed toward lower values (*i.e.,* a larger percentage of test items having a TER between 0 and 10).

On one side, a higher proportion of (near-)perfect test instances, requiring minimal or no corrections, increases the likelihood that APE systems will make unnecessary edits, which will be penalized by automatic evaluation metrics. Conversely, less skewed distributions may be easier to handle, as they provide automatic systems with more opportunities for improvement, with a larger number of test instances necessitating revision. In the lack of more focused analyses on this aspect, we can hypothesize that under ideal conditions from the APE standpoint, the peak of the distribution would correspond to "post-editable" translations containing enough errors that leave some margin for focused corrections but not too many errors to be so unintelligible to require a whole re-translation from scratch.[16] In light of the above observations, the APE 2023 test set can be considered as particularly challenging. As illustrated in Figure 2, the TER distribution exhibits a U-shaped (bimodal) pattern, characterized by two prominent peaks corresponding to the two most critical regions within the $0-100$ TER range. At one extreme, the first peak corresponds to the vast majority of test instances (about 45% of the total) that can be considered as perfect or near-perfect translations (*i.e.,* 0<TER<5), which implies a high chance of applying unnecessary corrections. At the other extreme, the second peak corresponds to a significant portion of test items (about 20%) that can be considered as too poor and unintelligible (*i.e.,* $95<TER<100$) to grant the safe application of any post-editing strategies. Although the remaining portion of the test set falls almost entirely in the range of "post-editable" outputs (*i.e.,* $10<TER<40$), its small size significantly reduces the potential for improvement through the APE process. Overall, this year's test set deviates significantly from all previous ones, where the TER distributions have never been characterized by such a pronounced bimodal pattern. In light of this, we can conclude that while, on the one hand, the repetition rate and machine translation quality do not provide sufficiently convincing insights to justify performance below the baseline for the official submissions, on the other hand, the TER distribution has posed a significant challenge for this year's participants.

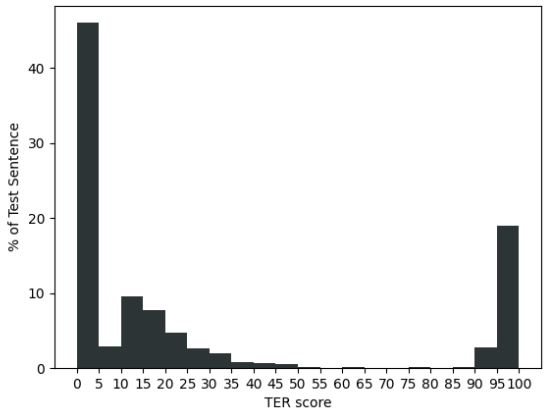


Figure 2: TER distribution in the APE 2023 English-Marathi test set.

5.2 Human Evaluation

We conducted a human evaluation of the primary system submissions to complement the automatic evaluations. However, this could be performed only for the official system submissions, as the late submission was received after the conclusion of the human assessments. This section discusses our evaluation procedure and the results obtained from it.

5.2.1 Evaluation Procedure

We provided annotation guidelines to professional translators who are native speakers of the target language. The same guidelines were also used to collect Indic language quality estimation shared task dataset (Zerva et al., 2022). The annotators provided a source-based direct assessment (DA) (Gra-

to 0.78.

[16]For instance, based on the empirical findings reported in (Turchi et al., 2013), TER=0.4 is the threshold that, for human post-editors, separates the "post-editable" translations from those that require complete rewriting from scratch.

	Avg DA	Avg z
test.pe	83.76	0.426
KU_UPs-filtered4-PRIMARY	66.56	−0.171
test.mt	65.86	−0.138
kaistai_prompt-w-cot_primary	64.79	−0.116

Table 6: Results for the human evaluation campaign for the En-Mr language pair. Systems ordered by DA score; systems within a cluster are considered tied; lines indicate clusters according to Wilcoxon rank-sum test p <0.05.

ham et al., 2013b; Cettolo et al., 2017; Bojar et al., 2018) score to each segment containing the source and the APE system output. We hired 4 translators to evaluate the two primary system submissions (KU_UP & KAISTAI), manually post-edited segments (*test.pe*), and the MT Output (*test.mt*). We chose to allocate an equal number of instances to each translator after shuffling, and only a single DA annotation was collected for each instance (Toral, 2020). Shuffling the instances before allocation helps prevent annotator bias towards a single system in the direct assessments.

The annotation guidelines provide a detailed description of potential adequacy and fluency-based errors based on which the translator could estimate the direct assessment score range. However, the translators were additionally instructed to prioritize adequacy errors and focus on assessing the semantic similarity between the source and the system output. The annotators manually entered the DA score between 0-100. The collected DA annotations were unshuffled based on the segment IDs, which were unknown to the translators. We expected the human post-editing to be of higher quality compared to APE system submissions and, consequently, better than the MT baseline.

5.2.2 Evaluation Results

We present the results obtained from the human evaluation campaign in Table 6. As expected, the human post-edited segments were rated the highest at 83.76 mean DA score. However, contrary to automatic evaluation, the submission by KU_UP was rated slightly better than the MT baseline (test.mt). But, the score difference in both cases-human and automatic evaluation, seems insignificant. Additionally, as per the Wilcoxon Rank-sum test, KU_UP and MT baseline score distributions seem to be in a cluster. In line with the automatic evaluation, the mean DA obtained by the submission from *kaistai* was rated the lowest at 64.79, lower than the MT baseline at 65.86. This submission utilizes LLMs to perform the APE task and raises a question on the viability of LLMs for APE when a low-resource language is concerned. LLMs are mostly fine-tuned and/or evaluated on task datasets in English (Hendrycks et al., 2020; Longpre et al., 2023), and there remain unanswered questions on their viability for complex and challenging multilingual tasks like APE. Owing to a challenging test set this year, our analysis highlights the difficulty posed by the task and implores us to consider a different setting in which the APE task can perhaps gain assistance through a translation quality signal. QE systems have been explored for assisting the APE task in a supervised multi-task scenario, which intuitively helps the model perform better at both tasks.

6 Conclusion

We presented the results from the 9^{th} shared task on Automatic Post-Editing at WMT. In continuity with the 2022 round, the task focused on the automatic correction of NMT outputs generated by a black-box English-Marathi system. The three participating systems were evaluated both automatically (with TER as the primary metric, BLEU, and ChrF) and manually. According to automatic evaluation results, only one (late) submission succeeded in outperforming the *do-nothing* baseline. The analysis of this year's data suggests that one of the main causes of difficulty might be the bimodal, U-shaped TER distribution of the test instances, which substantially differs from the test set distributions observed in all previous rounds (skewed but a pattern closer to normal). Our manual evaluation confirms the automatic evaluation outcomes and affirms the challenge posed by APE for the current approaches. We observe that one of the systems performs quite close to the MT baseline while the other performs well below the same. Additionally, the lack of multilingual datasets in LLM training/benchmarking raises a question on the viability of performing challenging multilingual tasks like APE. All in all, these findings advocate for further research on this challenging problem, which, far from being solved, this year revealed new nuances in terms of difficulty. Next year, we plan to introduce two new low-resource language pair datasets for the APE task. Future developments will also

likely include a re-definition of some aspects of the evaluation settings, which have remained relatively stable over the years. For instance, the set of automatic evaluation metrics will likely be reconsidered and expanded so as to include more semantics-oriented measures, with an eye on the advent of large language models increasingly adopted also for APE.

Acknowledgements

We would like to thank Zibanka Media Services Pvt. Ltd., and Techliebe, who worked with us to create the APE datasets for English-Marathi.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Aakash Banerjee, Aditya Jain, Shivam Mhaskar, Sourabh Dattatray Deoghare, Aman Sehgal, and Pushpak Bhattacharyya. 2021. Neural machine translation in low-resource setting: a case study in english-marathi pair. In *Proceedings of the 18th Biennial Machine Translation Summit (Volume 1: Research Track)*, pages 35–47.

Pushpak Bhattacharyya, Rajen Chatterjee, Markus Freitag, Diptesh Kanojia, Matteo Negri, and Marco Turchi. 2022. Findings of the WMT 2022 shared task on automatic post-editing. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 109–117, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 2–14, Tokyo, Japan. International Workshop on Spoken Language Translation.

Rajen Chatterjee, Christian Federmann, Matteo Negri, and Marco Turchi. 2019. Findings of the WMT 2019 shared task on automatic post-editing. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 11–28, Florence, Italy. Association for Computational Linguistics.

Rajen Chatterjee, Markus Freitag, Matteo Negri, and Marco Turchi. 2020. Findings of the WMT 2020 shared task on automatic post-editing. In *Proceedings of the Fifth Conference on Machine Translation*, pages 646–659, Online. Association for Computational Linguistics.

Rajen Chatterjee, Matteo Negri, Raphael Rubino, and Marco Turchi. 2018. Findings of the WMT 2018 shared task on automatic post-editing. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 710–725, Belgium, Brussels. Association for Computational Linguistics.

Rajen Chatterjee, Marion Weller, Matteo Negri, and Marco Turchi. 2015. Exploring the Planet of the APEs: a Comparative Study of State-of-the-art Methods for MT Automatic Post-Editing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 156–161, Beijing, China. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013a. Continuous Measurement Scales in Human Evaluation of Machine Translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013b. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Jyotsana Khatri, Rudra Murthy, Tamali Banerjee, and Pushpak Bhattacharyya. 2021. Simple measures of bridging lexical divergence help unsupervised neural machine translation for low-resource languages. *Machine Translation*, 35(4):711–744.

Xiaobo* Liang, Lijun* Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. 2021. R-drop: Regularized dropout for neural networks. In *Proceedings of NeurIPS*.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*.

Hyeonseok Moon, Seugnjun Lee, Chanjun Park, Jaehyung Seo, Sugyeong Eo, and Heuiseok Lim. 2023. What is the Resultful Data?: Empirical Study on the Adaptability of the Automatic Post-Editing Training Data. In *Proceedings of the Eighth Conference on Machine Translation (WMT23)*, Singapore.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, pages 223–231.

Antonio Toral. 2020. Reassessing claims of human parity and super-human performance in machine translation at WMT 2019. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 185–194, Lisboa, Portugal. European Association for Machine Translation.

Marco Turchi, Matteo Negri, and Marcello Federico. 2013. Coping with the subjectivity of human judgements in MT quality estimation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 240–251, Sofia, Bulgaria. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Jiawei Yu, Min Zhang, Yanqing Zhao, Xiaofeng Zhao, Yuang Li, Chang Su, Yinglu Li, Miaomiao Ma, Shimin Tao, and Hao Yang. 2023. HW-TSC's Participation in the WMT 2023 Automatic Post Editing Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT23)*, Singapore.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Findings of the WMT 2023 Shared Task on Low-Resource Indic Language Translation

Santanu Pal
Wipro AI Lab45

Partha Pakray
NIT Silchar

Sahinur Rahman Laskar
UPES, Dehradun

Lenin Laitonjam
NIT Mizoram

Vanlalmuansangi Khenglawt
Mizoram University

Sunita Warjri
MIT-WPU, Pune

Pankaj Kundan Dadure
UPES, Dehradun

Sandeep Kumar Dash
NIT Mizoram

Abstract

This paper presents the results of the low-resource Indic language translation task organized alongside the Eighth Conference on Machine Translation (WMT) 2023. In this task, participants were asked to build machine translation systems for any of four language pairs, namely, English-Assamese, English-Mizo, English-Khasi, and English-Manipuri. For this task, the IndicNE-Corp1.0 dataset is released, which consists of parallel and monolingual corpora for northeastern Indic languages such as Assamese, Mizo, Khasi, and Manipuri. The evaluation will be carried out using automatic evaluation metrics (BLEU, TER, RIBES, COMET, ChrF) and human evaluation.

1 Introduction

Low-resource Indic languages refer to the vast array of languages spoken in India, which, unfortunately, possess limited linguistic resources available for their study and development. These languages typically suffer from a combination of factors that set them apart from the more prominent and widely supported languages spoken in the country. The challenges these languages face include having a smaller number of speakers, a relative lack of governmental support, inadequate documentation, and limited access to technological resources.

India is renowned for its linguistic diversity, with a rich tapestry of languages spoken across the subcontinent. The Eighth Schedule of the Indian Constitution officially recognizes 22 languages, which receive significant government backing and protection. However, beyond these major languages, numerous smaller languages and dialects are spoken by various indigenous and minority communities throughout the country.

These low-resource Indic languages face a series of interconnected challenges that make their preservation and promotion difficult: lack of written scripts, limited vocabulary resources, inadequate linguistic research, and insufficient digital content. The collective impact of these factors makes it challenging to preserve and promote low-resource Indic languages. As a consequence, they are at risk of falling into disuse, with their speakers shifting to more widely recognized languages. Efforts to document, revitalize, and support these languages are crucial not only for linguistic diversity but also for the preservation of cultural heritage and the rights of minority language communities in India.

Efforts are being made by various organizations, researchers, and language enthusiasts to address the issues faced by low-resource Indic languages (Pal et al., 2013a,b; Pal, 2018). These initiatives involve language documentation, the development of writing systems, the creation of linguistic resources such as parallel corpora (Ramesh et al., 2022), parallel fragment extraction from comparable corpora (Gupta et al., 2013; Pal et al., 2014), dictionaries and grammars, and the promotion of language use through educational programs and digital platforms.

Technology indeed plays a pivotal role in supporting low-resource Indic languages. In recent years, machine learning and natural language processing techniques have been harnessed to create innovative solutions for these languages, including speech recognition, machine translation, and text-to-speech systems. These technological advancements offer a transformative potential in addressing the linguistic challenges faced by these languages and can have a profound impact on their preservation and revitalization.

To work towards the goal of supporting low-resource Indic languages, we organized the "Indic MT Shared Task" focusing on several less-popular languages that belong to different language families. These languages include As-

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 682–694

samese (Indo-Aryan), Mizo (Sino-Tibetan), Khasi (Austroasiatic), and Manipuri (Sino-Tibetan). In this shared task, we present IndicNE-Corp1.0 in which parallel (English-Assamese (en–as), English–Mizo (en–lus), English–Khasi (en–kha), English–Manipuri (en–mni)) and monolingual (Assamese, Mizo, Khasi, Manipuri) corpora for northeastern Indic languages available.

2 Shared Task: Low-Resource Indic Language Translation

In recent years, there has been significant improvement in the performance of machine translation (MT) systems. This progress can be attributed to the development of new techniques, such as multilingual translation and transfer learning. As a result, the benefits of MT are no longer restricted to users of widely spoken languages. This advancement has led to a growing interest within the research community in expanding MT coverage to encompass a wider range of languages, each with its unique geographical presence, degree of diffusion, and level of digitalization.

However, despite the enthusiasm for extending MT to more languages and users, there remains a substantial challenge. The challenge stems from the fact that MT methods typically require large volumes of parallel data for training high-quality translation systems. This requirement has proven to be a major hurdle, particularly when dealing with low-resource languages where obtaining such extensive parallel data can be exceedingly difficult. Consequently, there is a pressing need to develop MT systems that can perform well even when trained on relatively small parallel datasets. The ability to achieve effective machine translation with limited resources is of paramount importance for increasing accessibility and usability across a wide spectrum of languages and linguistic communities. In this translation task, our focus was on the following language pairs (both directions for each):

- Subtask-1 : English ↔ Assamese
- Subtask-2 : English ↔ Mizo
- Subtask-3 : English ↔ Khasi
- Subtask-4 : English ↔ Manipuri

In this translation task, participants had the opportunity to submit up to 1 PRIMARY system for each language pair/translation direction, where no additional parallel data was permitted for training. Additionally, participants could submit up to 2 CONTRASTIVE systems for each language pair/translation direction. This structure allowed participants to showcase their translation systems under various conditions and constraints, including the absence of additional parallel data in the case of PRIMARY systems.

3 Dataset: IndicNE-Corp1.0

In the creation of IndicNE-Corp1.0, we compiled datasets from our prior research projects, including contributions from Laskar et al. (2020, 2022); Khenglawt et al. (2022); Laskar et al. (2021); Laitonjam and Ranbir Singh (2021). These datasets served as the foundation for constructing both parallel and monolingual corpora. In our earlier works, we undertook the development of English-Assamese (eng-asm) (Laskar et al., 2020, 2022), English-Mizo (eng-lus) (Khenglawt et al., 2022), English-khasi (eng-kha) (Laskar et al., 2021), English-Manipuri (eng-mni) (Laitonjam and Ranbir Singh, 2021) parallel and monolingual corpora for Assamese, Mizo, Khasi and Manipuri languages. The different online sources were explored that include Bible, multilingual online dictionary (Xobdo and Glosbe), multilingual question paper, PMIndia[1] (Haddow and Kirefu, 2020), web pages, blogs and online news papers. The collected data statistics for parallel (train, validation and test set) and monolingual corpora are presented in subsequent sections below. For primary investigation in this shared task, we have not included very complex sentences in the test set.

3.1 Assamese

Assamese exhibits a subject-object-verb (SOV) word order, in contrast to the subject-verb-object (SVO) word order found in English. Additionally, it is characterized as an agglutinative language, as discussed by Sarma et al. (2017) and Baruah et al. (2021), signifying its propensity to incorporate suffixes and prefixes into words to convey diverse grammatical meanings. This intricacy poses a notable challenge for machine translation systems, as they must accurately analyze and generate these intricate word forms.

Furthermore, Assamese boasts a complex verb conjugation system encompassing tense, aspect, mood, and agreement markers. These markers hold

[1] http://data.statmt.org/pmindia/v1/ parallel/

the power to significantly alter the meaning of a verb, making it a formidable task for translation systems to capture these subtleties with precision. The data statistics for the English-Assamese parallel data are presented in Table 1.

Type	Sentences	Tokens	
		eng	asm
Train	50,000	969,623	825,063
Validation	2,000	31,503	25,929
Test	2,000	32,466	27,483

Table 1: English-Assamese parallel data statistics for train, valid, and test set

3.2 Mizo

Mizo follows the object-subject-verb order when the object is considered. Mizo is a tonal language (Lalrempuii et al., 2021; Khenglawt et al., 2022), which means that differences in pitch or tone can represent different meanings. The vowels (a, aw, e, i, o, u) primarily indicate intonation. In the Mizo language, the main tones are rising, falling, high, and low. For example, depending on the tone, the word "ban" in Mizo can mean a pillar, the arm, to stretch, arrive at, sticky, or dismiss. A circumflex (ˆ) is frequently used to indicate long intonations (primarily to distinguish them from short intonations). Mizo is an agglutinative and highly inflected language with declension of nouns and pronouns. It also has many monosyllables and decomposable polysyllables, with meaning derived from each syllable. A sentence's tense can be changed by including particles such as "ang," "dawn," "mek," "tawh," and so on. The data statistics for the English-Mizo parallel data are presented in Table 2.

Type	Sentences	Tokens	
		eng	lus
Train	50,000	981,468	1,06,2414
Validation	1,500	38,525	40,983
Test	2,000	21,905	25,098

Table 2: English-Mizo parallel data statistics for train, valid and test set

3.3 Khasi

Khasi follows the subject-verb-object word order. It orthography has 23 alphabets and has six vowels, the vowels are "a e i ï o u". In Khasi orthography the alphabets "c f q v x z" are not present and instead the letters "ï ñ ng" are present which makes the orthography to be different from English or other orthographies (Warjri et al., 2021). Khasi is rich in subject agreement markers. Subject agreement is indicated by verbs, adjectives, and adverbs. Nouns have their own grammatical number and gender. In morphology, Khasi is mostly isolating; while some words are derived through specific morphological processing, others are found standing alone with no morphology indicated. As a result, (a) word categories such as Nouns, Verbs, Adjectives, and so on are invariant, and (b) words are mostly mono-morphemic in nature, so it is common to encounter only isolating words in a single long sentence or discourse. The data statistics for the English-Khasi parallel data are presented in Table 3.

Type	Sentences	Tokens	
		eng	kha
Train	24,000	7,29,930	8,75,545
Validation	1,000	24,609	37,407
Test	1,000	24,150	35,901

Table 3: English-Khasi parallel data statistics for train, valid and test set

3.4 Manipuri

Manipuri language uses Bengali script[2] and Meetei mayek[3] in written form. In this dataset, we use the Bengali script. Manipuri language also has an extensive suffix with limited prefixation and verb-final word order in a sentence, i.e., subject-object-verb order (Huidrom et al., 2021). Linguistic features of this language include agglutinative verb morphology, tone, the absence of grammatical person, number, gender, and a prevalence of aspect over tense. The data statistics for the English-Manipuri parallel data are presented in Table 4.

Type	Sentences	Tokens	
		eng	mni
Train	21,687	390,730	330,319
Validation	1,000	16,905	14,469
Test	1,000	14,886	12,775

Table 4: English-Manipuri parallel data statistics for train, valid, and test set

Table 5 presents the statistics for parallel data length differences among the four language pairs. In Figure 1, we illustrate the overlapping tokens between the test set and the training and validation sets for these same four language pairs. In addition to the parallel data, we have also made available monolingual corpora for Assamese, Mizo, Khasi,

[2]http://unicode.org/charts/PDF/U0980.pdf
[3]http://unicode.org/charts/PDF/UABC0.pdf

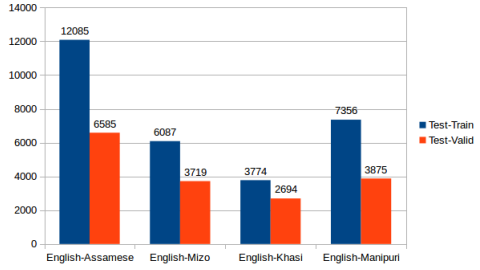


Figure 1: Overlapping tokens among test-train and test-validation data set of English-Assamese, English-Mizo, English-Khasi and English-Manipuri

and Manipuri. The monolingual data statistics per languages are presented in Table 6.

Language	Size (MB)	Sentences	Tokens
asm	805	2,624,715	49,232,154
lus	145	1,909,823	27,936,225
kha	104	182,737	22,140,361
mni	716	2,144,897	36,514,693

Table 6: Monolingual data statistics for Assamese, Mizo, Khasi and Manipuri languages

4 Participants and System Descriptions

In this shared task, a total of 31 teams registered and contributed, as indicated in Table 8, the released the dataset have been distributed among participants. In Table 7, we have compiled the system outputs submitted by participants, encompassing both primary and contrastive submission types.

Language Pair	Number of Participants
English-Assamese	13 (Primary), 11 (contrastive)
English-Mizo	10 (Primary), 8 (contrastive)
English-Khasi	11 (Primary), 8 (contrastive)
English-Manipuri	14 (Primary), 11 (contrastive)

Table 7: Number of participants in the low-resource Indic language translation task at WMT23

However, we have received system description papers from 9 teams and included concise system details for those teams where the authors provided such information.

CFILT-IITB (Gaikwad et al., 2023): The participant utilized phrase-pair injection (Sen et al., 2021), back-translation (Sennrich et al., 2016), and transfer learning with the help of large pre-trained multilingual IndicTrans2 model (Gala et al., 2023) to build NMT systems for the English-Assamese and English-Manipuri language pairs.

IOL Research (Zhang et al., 2023): The contributor used monolingual data to train two denoising language models similar to T5 (Raffel et al., 2020) and BART (Lewis et al., 2019), and then used parallel data to fine-tune the pre-trained language models to obtain two multilingual machine translation models. Besides, the multilingual machine translation models were used to translate English monolingual data into other multilingual data, forming multilingual parallel data as augmented data (Sennrich et al., 2016) to build NMT systems for English-Assamese, English-Mizo, English-Khasi, and English-Manipuri language pairs.

IACS-LRILT (Suman et al., 2023): The team IACS-LRILT used IndicBART (Dabre et al., 2022) pre-trained language model for fine-tuning the training data to build NMT systems for English-Assamese, and English-Manipuri language pairs.

GUIT-NLP (Ahmed et al., 2023): Team GUIT-NLP used back-translation (Sennrich et al., 2016) strategy and explored NMT systems by leveraging subword tokenization (Sennrich et al., 2015; Kudo and Richardson, 2018) and hyperparameters tuning for English-Assamese, English-Mizo, and English-Khasi language pairs.

NITS-CNLP (Singh et al., 2023): The NITS-CNLP team used the OpenNMT toolkit (Klein et al., 2017) and built a transformer-based (Vaswani et al., 2017) NMT model with hyperparameters tuning for the English-Manipuri language pair.

NICT-AI4B (Dabre et al., 2023): The group explored NMT systems by leveraging back-translation strategy (Sennrich et al., 2016) with denoising techniques (Lewis et al., 2020; Dabre et al., 2022) and fine-tuned IndicTrans2 model (Gala et al., 2023) for the English-Assamese, English-Mizo, English-Khasi, and English-Manipuri language pairs.

MUNI-NLP (Signoroni and Rychly, 2023): The participant explored transformer-based (Vaswani et al., 2017) NMT systems by investigating different hyperparameters tuning for English-Assamese, English-Mizo, English-Khasi, and English-Manipuri language pairs.

CUNI (Kvapilíková and Bojar, 2023): The CUNI team used back-translation (Sennrich et al., 2016) for data augmentation, denoising, leveraging multilingual masked language modelling, and

Data	Length	Number of Sentences			
		eng-asm	eng-lus	eng-kha	eng-mni
Test	1-10	435	1071	61	327
	11-20	1013	804	315	462
	21-30	481	120	381	164
	31-40	71	5	194	43
	41-50			49	4
Train	1-10	560	148	32	339
	11-20	910	437	341	335
	21-30	468	433	385	216
	31-40	62	292	194	86
	41-50		190	48	24
Valid	1-10	6895	10940	488	6351
	11-20	21032	16264	5559	7681
	21-30	18679	14316	7320	4764
	31-40	3245	8007	6656	1947
	41-50	149	473	3977	944

Table 5: Length-wise sentence group distribution for the test, train, and validation parallel data of English-Assamese, English-Mizo, English-Khasi, and English-Manipuri

built NMT systems for English-Assamese, English-Mizo, English-Khasi, and English-Manipuri language pairs.

ATULYA-NITS (Agrawal et al., 2023): This group used Google Colab, and trained the transformer model (Vaswani et al., 2017) using a T4 GPU for the English-Assamese, and English-Manipuri language pairs.

Organizer: The shared task organizer used the OpenNMT toolkit (Klein et al., 2017) and built biLSTM-based NMT systems with hyper-parameters tuning only on parallel data for the English-Assamese, English-Mizo, English-Khasi, and English-Manipuri language pairs.

5 Results and Discussion

We present results[4] for both directions of the four language pairs, namely, English-Assamese in Table 9, English-Mizo in Table 10, English-Khasi in Table 11, and English-Manipuri in Table 12. Here, we have reported the evaluation scores of those teams who submitted system output and their associated papers. To evaluate quantitative results, standard evaluation metrics (Papineni et al., 2002), namely, BLEU (bilingual evaluation under study), TER (translation error rate) (Snover et al., 2006), RIBES (rank-based intuitive bilingual evaluation score) (Isozaki et al., 2010), ChrF (character n-gram F-score) (Popović, 2015) and COMET (Rei et al., 2020). Moreover, we have hired linguistic experts who possess linguistic knowledge of the concerned language pair and randomly selected 20 sample sentences of primary submission type for manual evaluation (reported in Table 13 to 16). The human evaluator evaluates the candidate translations based on adequacy, fluency, and overall rating. Adequacy of translation measures the amount of meaning of reference translation, which is contained in a candidate translation. Furthermore, a translation is considered fluent if it is a well-formed sentence of the target language, irrespective of its correspondence with the reference translation. For example, given the reference translation to be "He wakes up early in the morning," the candidate translation "He is flying to Delhi" is inadequate, as it contains no content of the reference translation. However, the translation is fluent because the sentence has a proper meaning, and it is a well-formed sentence in the English language. The overall rating takes into account adequacy as well as fluency of candidate translation. An adequate and fluent translation is considered excellent and assigned a high overall rating. The human evaluation parameters have been rated on a scale of 0–5, with larger values signifying the better. Final adequacy, fluency, and overall rating scores are the average scores of individual test sentences.

[4] http://www2.statmt.org/wmt23/indic-mt-task.html

Team Name	Organization
BITS-P	Birla Institute of Technology and Science, Pilani, India
NITS-CNLP	**National Institute Of Technology, Silchar, India**
OneMT	IIIT-Hyderabad, India
SML lab	IISc, Bangalore, India
NICT-AI4B	**NICT Japan**
ANVITA	Centre for AI and Robotics (CAIR), India
MUNI-NLP	**Masaryk University, Czechia**
HV-NITS	National Institute Of Technology, Silchar, India
IREL-IIITH	IIIT HYDERABAD India
NVIDIA-India	NVIDIA, India
AIMLNLP-IITI	Indian Institute of Technology, Indore, India
NLP_NITH	NIT Hamirpur, India
TRANSSION MT	TRANNSION, China
CNLP-IISC	IISc, Bangalore, India
CUNI	**Charles University, Czechia**
A3-108	LTRC, IIIT Hyderabad, India
IOL Research	**Transn, China**
SLP-BV	Banasthali Vidyapith, India
IACS-LRILT	**Indian Assosciation for the Cultivation of Science, India**
NITR	NIT Rourkela, India
IIT-NLP lab	IIT dharwad, India
Team SiggyMorph	University of British Columbia, Canada
Lexical wizards	Kalinga Institute of Industrial Technology, India
JUNLP	Jadavpur University, India
ATULYA-NITS	**National Institute of Technology, Silchar, India**
CFILT-IITB	**Indian Institute of Technology, Bombay, India**
COGNITIVE LAB-IIITM	Indian Institute of Information Technology, Manipur, India
LRNMT-IIITH	IIIT Hyderabad, India
GUIT-NLP	**Gauhati University, India**
TRDDC	TCS Research, India
HW-TSC	Huawei Translate Center, China

Table 8: Registered participants in the low-resource Indic language translation task at WMT23 and dataset released to them. Not all the teams participated in all language pairs. **Bold marks** are those who submitted system outputs and system description papers

Discussion:

- For both directions of English-Assamese, Team: IACS-LRILT attains the best BLEU score (as shown in Table 9). They utilized the IndicBART language model in fine-tuning the training model. Also, Assamese-to-English translation attains higher scores than English-to-Assamese translation. It is due to the fact that Assamese is a highly inflectional, morphologically rich, and agglutinative language.
- For both directions of English-Mizo, Team: NICT-AI4B attains the best BLEU score (as shown in Table 10). They utilized IndicTrans2 model in fine-tuning the training model. It is observed that encountering tonal words for English-to-Mizo translation is a challenging task.
- For both directions of English-Khasi, Team: IOL Research attains the best BLEU score (as shown in Table 11). They used denoising language models (T5 / BART) and data augmentation techniques.
- For English-to-Manipuri translation, Team: CUNI attains the best BLEU score (as shown in Table 12). They used data augmentation, denoising, leveraging multilingual, and masked language modelling techniques. And, Manipuri-to-English translation, Team: IACS-LRILT attains the best BLEU score (as shown in Table 12) by utilizing the IndicBART language model in fine-tuning the training model. Also, it is observed that Manipuri-to-English translation attains higher scores than English-to-Manipuri translation. This is due to the fact that Manipuri is a morphologically rich and highly agglutinative language.
- In human evaluation, it is noticed that fluency scores are better than adequacy scores for all language pairs submission. The reason behind this is that NMT systems are well known for producing fluent translations (Koehn and Knowles, 2017).

6 Conclusion

We presented the results of the participating teams in the four language pairs translation task in terms of automatic and human evaluation metrics. We released a dataset, namely, IndicNE-Corp1.0 in the shared task on low-resource Indic language translation at the eighth conference on machine translation (WMT) 2023. The dataset comprises four low-resource languages, namely, Assamese, Mizo, Khasi, and Manipuri which belong to the northeastern region of India. In the future, we will include more northeastern Indic language datasets in addition to increasing the existing dataset size.

Comments

A few teams, namely, TRANSSION MT (TRANNSION, China), HW-TSC (Huawei Translate Center, China), ANVITA (Centre for AI and Robotics (CAIR), India), COGNITIVE LAB-IIITM (Indian Institute of Information Technology, Manipur, India) and NITR (NIT Rourkela, India) submitted system results but unfortunately did not submit the associated system description paper. Therefore, we have not reported their results in this paper.

Acknowledgements

We would like to thank all the participants for their active participation in this shared task. Also, thankful to the organizers and reviewers of the Eighth Conference on Machine Translation (WMT) 2023.

References

Goutam Agrawal, Rituraj Das, Anupam Biswas, and Dalton Meitei Thounaojam. 2023. Neural machine translation for english - manipuri and english - assamese.

Mazida Ahmed, Kuwali Talukdar, Parvez Boruah, Shikhar Kumar Sarma, and Kishore Kashyap. 2023. Guit-nlp's submission to shared task: Low resource indic language translation.

Rupjyoti Baruah, Rajesh Kumar Mundotiya, and Anil Kumar Singh. 2021. Low resource neural machine translation: Assamese to/from other indo-aryan (indic) languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(1).

Raj Dabre, Jay Gala, and Pranjal Chitale. 2023. Nict-ai4b's submission to the indic mt shared task in wmt 2023.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics.

Team Name	Translation Type	BLEU	ChrF	RIBES	TER	COMET
IACS-LRILT	English-To-Assamese (Primary)	**34.82**	56.58	0.87	55.10	0.77
	English-To-Assamese (Contrastive-1)	34.71	56.59	0.87	54.75	0.78
	English-To-Assamese (Contrastive-2)	6.57	39.71	0.45	86.26	0.79
	Assamese-To-English (Primary)	**66.36**	75.88	0.93	37.44	0.84
	Assamese-To-English (Contrastive-1)	66.33	75.88	0.93	37.38	0.84
	Assamese-To-English (Contrastive-2)	23.19	48.42	0.61	71.79	0.75
CFILT-IITB	English-To-Assamese (Primary)	18.15	50.16	0.53	75.53	0.80
	English-To-Assamese (Contrastive)	18.15	50.16	0.53	75.53	0.80
	Assamese-To-English (Primary)	35.24	57.73	0.70	60.85	0.80
	Assamese-To-English (Contrastive)	35.24	57.73	0.70	60.85	0.80
NICT-AI4B	English-To-Assamese (Primary)	17.03	45.31	0.58	76.57	0.78
	English-To-Assamese (Contrastive-2)	21.07	51.71	0.58	73.03	0.81
	English-To-Assamese (Contrastive-1)	18.09	51.98	0.57	73.41	0.82
	Assamese-To-English (Primary)	27.02	50.71	0.71	62.46	0.76
	Assamese-To-English (Contrastive-1)	37.28	59.97	0.72	58.81	0.81
	Assamese-To-English (Contrastive-2)	36.97	59.82	0.72	58.53	0.81
IOL Research	English-To-Assamese (Primary)	14.35	43.87	0.63	73.37	0.78
	English-To-Assamese (Contrastive)	14.10	43.66	0.63	72.77	0.78
	Assamese-To-English (Primary)	28.73	51.99	0.76	57.06	0.78
	Assamese-To-English (Contrastive)	27.83	51.45	0.76	57.44	0.78
CUNI	English-To-Assamese (Primary)	13.92	41.66	0.60	80.26	0.76
	English-To-Assamese (Contrastive-1)	3.98	41.57	0.59	78.91	0.75
	English-To-Assamese (Contrastive-2)	3.88	39.68	0.59	80.34	0.75
	Assamese-To-English (Primary)	20.71	44.94	0.69	73.56	0.72
	Assamese-To-English (Contrastive-1)	17.49	42.21	0.65	80.90	0.70
	Assamese-To-English (Contrastive-2)	16.85	41.55	0.65	85.24	0.70
Organizer	English-To-Assamese (Primary)	8.57	25.24	0.44	86.14	0.59
	Assamese-To-English (Primary)	11.28	28.70	0.53	83.10	0.56
MUNI-NLP	English-To-Assamese (Primary)	7.96	27.31	0.31	91.38	0.59
	Assamese-To-English (Primary)	11.29	30.13	0.64	73.39	0.64
ATULYA-NITS	English-To-Assamese (Primary)	5.47	21.66	0.32	96.76	0.57
	Assamese-To-English (Primary)	8.50	24.36	0.45	89.53	0.53
GUIT-NLP	English-To-Assamese (Primary)	4.89	25.16	0.46	87.21	0.61
	English-To-Assamese (Contrastive-1)	4.27	24.59	0.43	90.13	0.59
	English-To-Assamese (Contrastive-2)	3.75	22.65	0.42	93.57	0.58
	Assamese-To-English (Primary)	5.50	25.81	0.56	80.10	0.57
	Assamese-To-English (Contrastive-1)	4.70	24.96	0.55	81.53	0.56
	Assamese-To-English (Contrastive-2)	4.14	23.73	0.53	83.41	0.55

Table 9: Automatic evaluation scores of participated teams for English-Assamese language pair

Team Name	Translation Type	BLEU	ChrF	RIBES	TER	COMET
NICT-AI4B	English-To-Mizo (Primary)	33.18	56.73	0.73	55.68	0.70
	English-To-Mizo (Contrastive-2)	**33.64**	56.88	0.72	57.71	0.71
	English-To-Mizo (Contrastive-1)	26.47	50.60	0.66	65.97	0.69
	Mizo-To-English (Primary)	32.47	51.33	0.69	60.56	0.67
	Mizo-To-English (Contrastive-2)	**33.30**	52.74	0.70	60.87	0.68
	Mizo-To-English (Contrastive-1)	28.47	47.93	0.61	67.54	0.69
CUNI	English-To-Mizo (Primary)	31.20	54.56	0.76	54.54	0.70
	English-To-Mizo (Contrastive-1)	31.28	54.58	0.76	54.20	0.70
	English-To-Mizo (Contrastive-2)	30.66	54.48	0.76	54.98	0.69
	Mizo-To-English (Primary)	29.47	49.98	0.73	60.44	0.66
	Mizo-To-English (Contrastive-1)	28.63	48.58	0.72	62.21	0.65
	Mizo-To-English (Contrastive-2)	28.53	49.51	0.73	62.55	0.66
IOL Research	English-To-Mizo (Primary)	28.24	54.02	0.78	53.04	0.70
	English-To-Mizo (Contrastive-1)	27.74	53.71	0.78	53.40	0.70
	Mizo-To-English (Primary)	32.54	51.83	0.78	53.48	0.71
	Mizo-To-English (Contrastive-1)	31.37	50.94	0.77	55.37	0.70
Organizer	English-To-Mizo (Primary)	23.67	45.1	0.71	62.29	0.67
	Mizo-To-English (Primary)	22.59	39.53	0.66	68.83	0.57
GUIT-NLP	English-To-Mizo (Primary)	23.29	46.72	0.75	59.93	0.68
	English-To-Mizo (Contrastive-1)	23.78	48.06	0.75	58.07	0.69
	Mizo-To-English (Primary)	18.81	40.33	0.66	73.65	0.57
	Mizo-To-English (Contrastive-1)	18.51	41.32	0.67	73.70	0.60
MUNI-NLP	English-To-Mizo (Primary)	20.48	45.60	0.73	61.22	0.68
	Mizo-To-English (Primary)	23.16	43.02	0.72	62.31	0.63

Table 10: Automatic evaluation scores of participated teams for English-Mizo language pair

Team Name	Translation Type	BLEU	ChrF	RIBES	TER	COMET
IOL Research	English-To-Khasi (Primary)	**21.63**	44.47	0.72	62.10	0.68
	English-To-Khasi (Contrastive)	21.48	44.30	0.65	62.55	0.68
	Khasi-To-English (Primary)	**20.72**	43.34	0.72	71.78	0.63
	Khasi-To-English (Contrastive)	20.60	43.09	0.58	71.35	0.63
NICT-AI4B	English-To-Khasi (Primary)	19.95	43.30	0.68	66.47	0.67
	English-To-Khasi (Contrastive-2)	21.05	46.06	0.65	73.80	0.68
	English-To-Khasi (Contrastive-1)	20.77	43.82	0.65	69.51	0.68
	Khasi-To-English (Primary)	17.80	39.22	0.66	74.10	0.60
	Khasi-To-English (Contrastive-1)	20.06	40.33	0.58	78.44	0.60
	Khasi-To-English (Contrastive-2)	20.02	39.82	0.59	77.50	0.59
CUNI	English-To-Khasi (Primary)	16.64	39.92	0.65	70.69	0.67
	English-To-Khasi (Contrastive-1)	16.49	40.00	0.65	69.92	0.67
	English-To-Khasi (Contrastive-2)	15.79	38.79	0.65	71.29	0.66
	Khasi-To-English (Primary)	13.84	37.05	0.65	79.73	0.58
	Khasi-To-English (Contrastive-1)	12.71	36.32	0.66	81.37	0.57
	Khasi-To-English (Contrastive-2)	11.55	35.62	0.64	87.54	0.56
MUNI-NLP	English-To-Khasi (Primary)	13.90	37.31	0.61	73.99	0.65
	Khasi-To-English (Primary)	12.71	34.55	0.65	78.15	0.56
GUIT-NLP	English-To-Khasi (Primary)	10.41	33.31	0.63	71.67	0.64
	Khasi-To-English (Primary)	8.74	30.54	0.63	79.64	0.52
Organizer	English-To-Khasi (Primary)	10.08	31.13	0.59	75.57	0.62
	Khasi-To-English (Primary)	8.02	28.04	0.56	86.94	0.49

Table 11: Automatic evaluation scores of participated teams for English-Khasi language pair

Team Name	Translation Type	BLEU	ChrF	RIBES	TER	COMET
CUNI	English-To-Manipuri (Primary)	**29.50**	59.85	0.73	60.60	0.74
	English-To-Manipuri (Contrastive-1)	5.96	60.96	0.75	58.97	0.75
	English-To-Manipuri (Contrastive-2)	5.86	60.13	0.73	60.25	0.74
	Manipuri-To-English (Primary)	36.08	62.29	0.76	61.19	0.76
	Manipuri-To-English (Contrastive-1)	33.62	60.29	0.75	65.96	0.75
	Manipuri-To-English (Contrastive-2)	31.03	59.08	0.74	77.42	0.74
NICT-AI4B	English-To-Manipuri (Primary)	27.36	61.60	0.74	58.28	0.76
	English-To-Manipuri (Contrastive-2)	27.40	61.55	0.74	58.16	0.76
	English-To-Manipuri (Contrastive-1)	24.17	62.95	0.70	62.85	0.76
	Manipuri-To-English (Primary)	39.40	64.70	0.77	51.27	0.79
	Manipuri-To-English (Contrastive-1)	46.06	69.96	0.80	47.44	0.83
	Manipuri-To-English (Contrastive-2)	43.35	69.27	0.80	47.43	0.82
CFILT-IITB	English-To-Manipuri (Primary)	26.36	63.48	0.70	62.04	0.76
	English-To-Manipuri (Contrastive-1)	26.36	63.48	0.70	62.04	0.76
	Manipuri-To-English (Primary)	47.54	70.41	0.81	47.17	0.83
	Manipuri-To-English (Contrastive-1)	47.54	70.41	0.81	47.17	0.83
IACS-LRILT	English-To-Manipuri (Primary)	25.78	49.94	0.84	60.43	0.71
	English-To-Manipuri (Contrastive-1)	25.82	49.93	0.84	60.57	0.71
	English-To-Manipuri (Contrastive-2)	9.69	40.45	0.54	81.18	0.67
	Manipuri-To-English (Primary)	**69.75**	78.16	0.94	32.08	0.84
	Manipuri-To-English (Contrastive-1)	69.75	78.16	0.94	32.10	0.84
	Manipuri-To-English (Contrastive-2)	22.10	48.03	0.63	72.19	0.70
IOL Research	English-To-Manipuri (Primary)	23.51	60.03	0.74	60.68	0.75
	English-To-Manipuri (Contrastive)	23.05	59.85	0.70	61.04	0.75
	Manipuri-To-English (Primary)	42.68	67.55	0.83	46.27	0.82
	Manipuri-To-English (Contrastive)	42.48	67.51	0.80	46.31	0.82
NITS-CNLP	English-To-Manipuri (Primary)	22.75	48.35	0.61	70.02	0.70
	Manipuri-To-English (Primary)	26.92	48.64	0.65	67.62	0.66
Organizer	English-To-Manipuri (Primary)	21.58	45.97	0.61	69.76	0.69
	Manipuri-To-English (Primary)	24.86	46.37	0.64	70.26	0.63
MUNI-NLP	English-To-Manipuri (Primary)	19.65	53.26	0.66	69.70	0.72
	Manipuri-To-English (Primary)	32.18	58.71	0.76	56.35	0.74
	Manipuri-To-English (Contrastive)	32.18	58.71	0.74	67.86	0.74
ATULYA-NITS	English-To-Manipuri (Primary)	15.02	35.96	0.46	85.96	0.65
	Manipuri-To-English (Primary)	18.70	38.49	0.54	81.02	0.59

Table 12: Automatic evaluation scores of participated teams for English-Manipuri language pair

Team Name	Translation Type	Adequacy	Fluency	Overall Rating
NICT-AI4B	English-To-Assamese (Primary)	3.60	4.35	3.98
	Assamese-To-English (Primary)	3.75	4.30	4.03
CFILT-IITB	English-To-Assamese (Primary)	2.80	3.85	3.33
	Assamese-To-English (Primary)	3.50	4.35	3.93
IACS-LRILT	English-To-Assamese (Primary)	2.55	3.20	2.88
	Assamese-To-English (Primary)	3.20	3.35	3.28
IOL Research	English-To-Assamese (Primary)	3.10	4.20	3.65
	Assamese-To-English (Primary)	3.70	4.60	4.15
CUNI	English-To-Assamese (Primary)	3.60	4.05	3.82
	Assamese-To-English (Primary)	2.85	3.80	3.32
Organizer	English-To-Assamese (Primary)	1.60	3.05	4.65
	Assamese-To-English (Primary)	1.50	2.55	2.02
MUNI-NLP	English-To-Assamese (Primary)	1.35	3.35	2.35
	Assamese-To-English (Primary)	1.50	2.45	1.97
ATULYA-NITS	English-To-Assamese (Primary)	1.50	2.95	2.22
	Assamese-To-English (Primary)	1.30	2.60	3.90
GUIT-NLP	English-To-Assamese (Primary)	1.35	3.05	2.20
	Assamese-To-English (Primary)	1.00	2.45	3.45

Table 13: Human evaluation score of English-Assamese

Team Name	Translation Type	Adequacy	Fluency	Overall Rating
NICT-AI4B	English-To-Mizo (Primary)	3.60	4.25	3.92
	Mizo-To-English (Primary)	3.10	4.50	3.80
CUNI	English-To-Mizo (Primary)	2.85	4.35	3.60
	Mizo-To-English (Primary)	3.30	4.40	3.85
IOL Research	English-To-Mizo (Primary)	3.95	4.45	4.20
	Mizo-To-English (Primary)	3.75	4.55	4.15
Organizer	English-To-Mizo (Primary)	2.05	3.55	2.80
	Mizo-To-English (Primary)	1.60	3.35	2.47
MUNI-NLP	English-To-Mizo (Primary)	3.05	3.85	3.45
	Mizo-To-English (Primary)	2.50	4.20	3.35
GUIT-NLP	English-To-Mizo (Primary)	3.25	4.15	3.70
	Mizo-To-English (Primary)	2.00	3.75	2.87

Table 14: Human evaluation score of English-Mizo

Team Name	Translation Type	Adequacy	Fluency	Overall Rating
IOL Research	English-To-Khasi (Primary)	4.45	4.75	4.60
	Khasi-To-English (Primary)	4.30	4.70	4.50
NICT-AI4B	English-To-Khasi (Primary)	4.20	4.60	4.40
	Khasi-To-English (Primary)	3.70	4.40	4.05
CUNI	English-To-Khasi (Primary)	3.30	4.20	3.75
	Khasi-To-English (Primary)	3.40	4.40	3.90
MUNI-NLP	English-To-Khasi (Primary)	2.70	4.50	3.60
	Khasi-To-English (Primary)	2.65	4.05	3.35
GUIT-NLP	English-To-Khasi (Primary)	2.80	4.60	3.70
	Khasi-To-English (Primary)	2.45	3.80	3.12
Organizer	English-To-Khasi (Primary)	1.95	4.05	3.00
	Khasi-To-English (Primary)	1.80	3.45	2.62

Table 15: Human evaluation score of English-Khasi

Team Name	Translation Type	Adequacy	Fluency	Overall Rating
CUNI	English-To-Manipuri (Primary)	3.25	3.55	3.45
	Manipuri-To-English (Primary)	3.05	3.15	3.00
NICT-AI4B	English-To-Manipuri (Primary)	2.95	4.10	3.50
	Manipuri-To-English (Primary)	3.50	3.50	3.45
CFILT-IITB	English-To-Manipuri (Primary)	4.25	4.50	4.35
	Manipuri-To-English (Primary)	4.80	4.75	4.75
IACS-LRILT	English-To-Manipuri (Primary)	2.45	2.65	2.45
	Manipuri-To-English (Primary)	3.45	3.45	3.45
IOL Research	English-To-Manipuri (Primary)	2.80	4.60	3.70
	Manipuri-To-English (Primary)	3.95	4.00	3.95
NITS-CNLP	English-To-Manipuri (Primary)	2.45	3.05	2.70
	Manipuri-To-English (Primary)	2.15	2.50	2.20
Organizer	English-To-Manipuri (Primary)	2.50	3.50	2.95
	Manipuri-To-English (Primary)	2.05	2.10	2.05
MUNI-NLP	English-To-Manipuri (Primary)	3.00	3.50	3.15
	Manipuri-To-English (Primary)	3.20	3.30	3.20
ATULYA-NITS	English-To-Manipuri (Primary)	1.75	2.15	1.95
	Manipuri-To-English (Primary)	1.80	1.85	1.80

Table 16: Human evaluation score of English-Manipuri

Pranav Gaikwad, Meet Doshi, Sourabh Deoghare, and Pushpak Bhattacharyya. 2023. Machine translation advancements for low-resource indian languages in wmt23: Cfilt-iitb's effort for bridging the gap.

Jay Gala, Pranjal A Chitale, Raghavan AK, Sumanth Doddapaneni, Varun Gumma, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, et al. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.

Rajdeep Gupta, Santanu Pal, and Sivaji Bandyopadhyay. 2013. Improving mt system using extracted parallel fragments of text from comparable corpora. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 69–76.

Barry Haddow and Faheem Kirefu. 2020. Pmindia - A collection of parallel corpora of languages of india. *CoRR*, abs/2001.09907.

Rudali Huidrom, Yves Lepage, and Khogendra Khomdram. 2021. EM corpus: a comparable corpus for a less-resourced language pair Manipuri-English. In *Proceedings of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021)*, pages 60–67, Online (Virtual Mode). INCOMA Ltd.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952, Cambridge, MA. Association for Computational Linguistics.

Vanlalmuansangi Khenglawt, Sahinur Rahman Laskar, Santanu Pal, Partha Pakray, and Ajoy Kumar Khan. 2022. Language resource building and English-to-mizo neural machine translation encountering tonal words. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 48–54, Marseille, France. European Language Resources Association.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Ivana Kvapilíková and Ondřej Bojar. 2023. Low-resource machine translation systems for indic languages.

Lenin Laitonjam and Sanasam Ranbir Singh. 2021. Manipuri-English machine translation using comparable corpus. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 78–88, Virtual. Association for Machine Translation in the Americas.

Candy Lalrempuii, Badal Soni, and Partha Pakray. 2021. An improved english-to-mizo neural machine translation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 20(4).

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji Darsh Kaushik, Partha Pakray, and Sivaji Bandyopadhyay. 2021. EnKhCorp1.0: An English–Khasi corpus. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages*

(LoResMT2021), pages 89–95, Virtual. Association for Machine Translation in the Americas.

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, and Sivaji Bandyopadhyay. 2020. EnAsCorp1.0: English-Assamese corpus. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 62–68, Suzhou, China. Association for Computational Linguistics.

Sahinur Rahman Laskar, Riyanka Manna, Partha Pakray, and Sivaji Bandyopadhyay. 2022. A domain specific parallel corpus and enhanced english-assamese neural machine translation. *Computación y Sistemas*, 26(4):1669—-1687.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Santanu Pal. 2018. A hybrid machine translation framework for an improved translation workflow.

Santanu Pal, Mahammed Hasanuzzaman, Sudip Kumar Naskar, and Sivaji Bandyopadhyay. 2013a. Impact of linguistically motivated shallow phrases in pb-smt. In *ICON 2013*. https://www.researchgate.net/publication ....

Santanu Pal, Sudip Kumar Naskar, and Sivaji Bandyopadhyay. 2013b. Mwe alignment in phrase based statistical machine translation. In *Proceedings of the XIV Machine Translation Summit*, pages 61–68.

Santanu Pal, Partha Pakray, and Sudip Kumar Naskar. 2014. Automatic building and using parallel resources for smt from comparable corpora. In *Proceedings of the 3rd Workshop on Hybrid Approaches to Machine Translation (HyTra)*, pages 48–57.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Himangshu Sarma, Navanath Saharia, and Utpal Sharma. 2017. Development and analysis of speech recognition systems for assamese language using htk. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 17(1).

Sukanta Sen, Mohammed Hasanuzzaman, Asif Ekbal, Pushpak Bhattacharyya, and Andy Way. 2021. Neural machine translation of low-resource languages using SMT phrase pair injection. *Nat. Lang. Eng.*, 27(3):271–292.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Edoardo Signoroni and Pavel Rychly. 2023. Muni-nlp systems for low-resource indic machine translation.

Kshetrimayum Boynao Singh, Avichandra Singh Ningthoujam, Loitongbam Sanayai Meetei, Sivaji Bandyopadhyay, and Thoudam Doren Singh. 2023. Nits-cnlp low-resource neural machine translation systems of english-manipuri language pair.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Dhairya Suman, Atanu Mandal, Santanu Pal, and Sudip Naskar. 2023. Iacs-lrilt: Machine translation for low-resource indic languages.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Sunita Warjri, Partha Pakray, Saralin A. Lyngdoh, and Arnab Kumar Maji. 2021. Part-of-speech (pos) tagging using deep learning-based approaches on the designed khasi pos corpus. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(3).

Wenbo Zhang, Zeyu Yan, Qiaobo Deng, Jie Cai, and Hongbao Mao. 2023. Iol research machine translation systems for wmt23 low-resource indic language translation shared task.

ACES: Translation Accuracy Challenge Sets at WMT 2023

Chantal Amrhein[1,2*] and **Nikita Moghe**[3*] and **Liane Guillou**[4*]
[1]Textshuttle, Zurich
[2]Department of Computational Linguistics, University of Zurich
[3]School of Informatics, University of Edinburgh
[4]Department of Computer Science, RISE Research Institutes of Sweden
amrhein@textshuttle.com, nikita.moghe@ed.ac.uk, liane.guillou@ri.se

Abstract

We benchmark the performance of segment-level metrics submitted to WMT 2023 using the ACES Challenge Set (Amrhein et al., 2022). The challenge set consists of 36K examples representing challenges from 68 phenomena and covering 146 language pairs. The phenomena range from simple perturbations at the word/character level to more complex errors based on discourse and real-world knowledge. For each metric, we provide a detailed profile of performance over a range of error categories as well as an overall *ACES*-Score for quick comparison. We also measure the incremental performance of the metrics submitted to both WMT 2023 and 2022. We find that 1) there is no clear *winner* among the metrics submitted to WMT 2023, and 2) performance change between the 2023 and 2022 versions of the metrics is highly variable. Our recommendations are similar to those from WMT 2022. Metric developers should focus on: building ensembles of metrics from different design families, developing metrics that pay more attention to the source and rely less on surface-level overlap, and carefully determining the influence of multilingual embeddings on MT evaluation.

1 Introduction

Challenge sets are a useful tool in measuring the performance of systems or metrics on one or more specific phenomena of interest. They may be used to compare the performance of a range of *different* systems or to identify performance improvement/degradation between successive iterations of the *same* system.

Challenge sets exist for a range of natural language processing (NLP) tasks including Sentiment Analysis (Li et al., 2017; Mahler et al., 2017; Staliūnaitė and Bonfil, 2017), Natural Language Inference (McCoy and Linzen, 2019; Rocchietti et al., 2021), Question Answering (Ravichander et al., 2021), Machine Reading Comprehension (Khashabi et al., 2018), Machine Translation (MT) (King and Falkedal, 1990; Isabelle et al., 2017), and the more specific task of pronoun translation in MT (Guillou and Hardmeier, 2016).

The WMT 2021 Metrics shared task (Freitag et al., 2021) introduced the task of constructing contrastive challenge sets for the evaluation of MT metrics. Contrastive challenge sets aim to assess how well a given metric is able to discriminate between a *good* and *incorrect* translation of the *source* text. The provision of a *reference* translation allows for flexibility: it may be included for the assessment of reference-based (i.e. MT) metrics, or excluded for the assessment of reference-free (i.e. Quality Estimation (QE)) metrics.

We re-submitted ACES[1] (Amrhein et al., 2022), originally developed for the WMT 2022 challenge sets shared task (Freitag et al., 2022), to the corresponding shared task at WMT 2023. ACES largely focuses on translation accuracy errors and consists of 68 phenomena ranging from simple perturbations at the word/character level to more complex errors based on discourse and real-world knowledge. We report on both the performance of metrics submitted to WMT 2023, and on the incremental performance of those metrics that were submitted to both WMT 2022 and WMT 2023. We also repeat the analyses in Amrhein et al. (2022) for the WMT 2023 metrics to confirm whether the findings from WMT 2022 still hold.

Overall, we find similar trends to those observed last year. Again, we do not find one clear winner and whilst neural metrics tend to perform better than their non-neural counterparts, different categories of metrics exhibit different strengths and weaknesses. The major challenges identified in Amrhein et al. (2022) still hold: (i) reference-based metrics are still overly reliant on the reference

*Equal contribution by all authors.

[1]The ACES dataset is available at https://huggingface.co/datasets/nikitam/ACES

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 695–712
December 6–7, 2023.

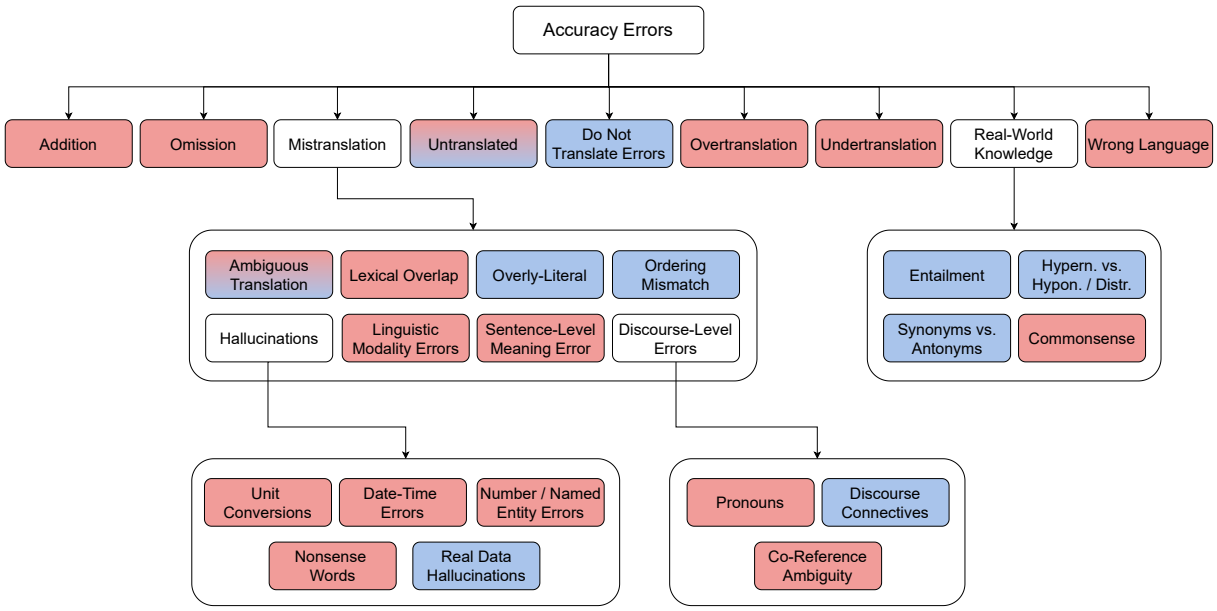


Figure 1: Diagram of the error categories on which our collection of challenge sets is based. Red means challenge sets are created automatically, and blue means challenge sets are created manually.

and do not pay enough attention to the source, (ii) reference-based metrics still rely on surface-level overlap, and (iii) the over-reliance on multilingual embeddings still persists – evidence from our analyses suggests that language agnostic representations present in the multilingual space may harm performance. Accordingly, our recommendations are also similar to those of last year. Metric developers should focus on: constructing ensembles of metrics with different design principles, developing metrics that also focus on information in the source, reducing dependence on surface-level overlap with the reference, and reassessing the impact of multilingual embeddings on MT evaluation.

With respect to incremental performance changes between metrics submitted to both 2022 and 2023, we observe mixed results. Whilst improvements are observed for some metrics, there is a degradation in performance for other metrics. However, even for those metrics for which an overall improvement was observed, this improvement was inconsistent across the top-level categories in ACES. Further, the performance even degraded for some categories.

2 ACES Overview

The Translation **A**ccuracy **C**hallen**E** Set (ACES) consists of 36,476 examples covering 146 language pairs and representing challenges from 68 linguistic phenomena. These phenomena are grouped into ten top-level categories: addition, omission, mistranslation, untranslated, do not translate errors, overtranslation, undertranslation, real-world knowledge, wrong language, and punctuation [2]. The mistranslation and real-world knowledge categories are further sub-categorised to include additional fine-grained categories. We illustrate the broad accuracy error categories in Fig 1 and give examples for each of the top-level categories in Appendix A.

The focus of ACES is on translation accuracy errors, reflecting the need to evaluate contemporary MT systems that are capable of producing fluent but potentially error-prone output. The selection of the top-level categories in the ACES error hierarchy is based on the *Accuracy* class in the Multidimensional Quality Metrics (MQM) ontology (Lommel et al., 2014), and extended to include translations defying *real-world knowledge* and translations in the *wrong language*. ACES includes a wide range of phenomena ranging from simple perturbations that involve the omission/addition of characters or tokens, to more complex scenarios involving mistranslations e.g. ambiguity or hallucinations in translation, untranslated elements of a sentence, discourse-level phenomena, and real-world knowledge.

Each ACES example consists of a *source* sentence, a *reference* translation, a *phenomenon label* indicating the error type, and two translation

[2]Note that although the focus of ACES is on *accuracy* errors, we also include a small set of *fluency* errors for the punctuation category.

	Mistranslation - Overly literal (Idioms)
SRC (de):	Er hat versucht, mir die Spielregeln zu erklären, aber **ich verstand nur Bahnhof**.
REF (en):	He tried to explain the rules of the game to me, but **I did not understand them**.
✓:	He tried to explain the rules of the game to me, but **it was all Greek to me**.
✗:	He tried to explain the rules of the game to me, but **I only understood train station**.
	Real-world Knowledge - Commonsense
SRC (en):	Die Luft im Haus war kühler als in der Wohnung.
REF (de):	The air in the house was cooler than in the apartment.
✓:	The air in the house was cooler than in the apartment because **the apartment** had a broken air conditioner.
✗:	The air in the house was cooler than in the apartment because **the house** had a broken air conditioner.

Table 1: Examples from the Mistranslation and Real-world Knowledge categories in ACES. An example consists of a source sentence (SRC), reference (REF), good (✓) and incorrect (✗) translations, and a phenomenon label indicating the error type. en: English, de: German. Top: the German idiom "ich verstand nur Bahnhof" has been translated in an overly-literal way in the incorrect translation. Bottom: the incorrect translation contains an error in commonsense reasoning as to why the air in the house was cooler than in the apartment.

hypotheses: an *incorrect* translation containing an error relating to the phenomenon of interest, and a *good* translation. Several examples from ACES are presented in Table 1. In the top example, from the *Mistranslation* error category, the incorrect translation contains an *overly literally* translation of the German *idiom* "ich verstand nur Bahnhof" (corresponding to the English expression "it was all Greek to me"). In the bottom example, from the *Real-world Knowledge* error category, both the good and incorrect translations contain additional information not present in the source sentence, however, the incorrect translation contains an error in *commonsense* reasoning as to why the air in the house was cooler than the apartment.

ACES examples were constructed from pre-existing datasets, using a range of automatic, semi-automatic, and manual methods. A detailed description of each of the phenomena in ACES can be found in Amrhein et al. (2022).

3 Related Work

Challenge sets have been used for several tasks (Li et al. (2017); McCoy and Linzen (2019); Ravichander et al. (2021), *inter alia*) to investigate the behaviour of these tasks under a specific phenomenon rather than the standard test distribution (Popović and Castilho, 2019). Lately, with the success of neural metrics, the development of challenge sets for MT evaluation has promoted great interest in studying the strengths and weaknesses of these metrics. We summarise here recent work on challenge sets for MT metric evaluation.

DEMETR (Karpinska et al., 2022), which comprises 31K English examples translated from ten languages, was developed for evaluating MT metric sensitivity to a range of 35 different types of linguistic perturbations, belonging to semantic, syntactic, and morphological error categories. These were divided into minor, major, and critical errors according to the type of perturbation, similar to the grading of error categories to compute the weighted ACES-Score. As in ACES, example generation was carefully designed to form minimal pairs such that the perturbed translation only differs from the actual translation in one aspect. The application of DEMETR in evaluating a suite of baseline metrics revealed a similar pattern to the analyses in Amrhein et al. (2022) - that metric performance varies considerably across the different error categories, often with no clear *winner*. It is worth noting that DEMETR and ACES each have their respective advantages: all examples in DEMETR have been verified by human annotators; ACES provides broader coverage in terms of both languages and linguistic phenomena.

In addition to ACES, three other datasets were submitted to the WMT 2022 challenge sets shared task (Freitag et al., 2022): SMAUG (Alves et al., 2022), the HWTSC challenge set (Chen et al., 2022), and the DFKI challenge set (Avramidis and Macketanz, 2022). These datasets differ from ACES in terms of their size, and the languages and phenomena/categories they cover (see Table 2).

Both SMAUG and HWTSC contained five different phenomena each pertaining to a single category of critical error for meaning change. In comparison, the DFKI challenge set has over 100 linguistically motivated phenomena, organised into 14 categories. Whereas the aim of ACES was to provide a broad coverage of language pairs, the other datasets provide an in-depth focus on specific lan-

	Ex.	Lang. pairs	Phenomena	Categories
SMAUG	632	2	5	5
HWTSC	721	1	5	5
DFKI	19,347	1	>100	14
ACES	36,476	146	68	10

Table 2: Comparison of challenge sets for MT metric evaluation in terms of: **Ex**amples, **Lang**uage**-pairs**, Phenomena, and Categories.

guage pairs: SMAUG (pt↔en and es→en), DFKI (de↔en), and HWTSC (zh↔en). Whilst there is a clear overlap between the ACES phenomena and those in SMAUG and HWTSC, many of the phenomena in the DFKI dataset are complementary such that in the case of evaluating metrics for the German-English pair, metric developers might consider benchmarking on both datasets.

4 Metrics

We list below the metrics that participated in the 2023 challenge set shared task and the baselines provided by the organisers.

4.1 Baseline Metrics

BERTScore (Zhang et al., 2020) uses contextual embeddings from pre-trained language models to compute the cosine similarity between the tokens in the hypothesis and the reference translation. The resulting similarity matrix is used to compute precision, recall, and F1-scores.

BLEURT-20 (Sellam et al., 2020) is a BERT-based (Devlin et al., 2019) regression model, which is first trained on scores produced by automatic metrics/similarity of pairs of reference sentences and their corrupted counterparts. It is then fine-tuned on WMT human evaluation data to provide a score for a hypothesis given a reference translation.

BLEU (Papineni et al., 2002) compares the token-level n-grams in the hypothesis with those in the reference translation. It then computes a precision score weighted by a brevity penalty.

chrF (Popović, 2017) provides a character n-gram F-score by computing overlaps between the hypothesis and reference translation.

COMET-22 (Rei et al., 2022) is an ensemble between a vanilla COMET model (Rei et al., 2020) trained with Direct Assessment (DA) scores and a multitask model that is trained on regression (MQM regression) and sequence tagging (OK/BAD word identification from MQM span annotations). These models are ensembled together using a hyperparameter search that weights different features extracted from these two evaluation models and combines them into a single score. The vanilla COMET model is trained with DAs ranging from 2017 to 2020 while the Multitask model is trained using DAs ranging from 2017 to 2020 plus MQM annotations from 2020 (except for en-ru which uses TedTalk annotations from 2021).

COMET-Kiwi (Rei et al., 2022) ensembles two QE models similarly to COMET-22. The first model follows the classic Predictor-Estimator QE architecture where MT and source are encoded together. This model is trained on DAs ranging from 2017 to 2019 and then fine-tuned on DAs from MLQE-PE (the official DA from the QE shared task). The second model is the same multitask model used in the COMET-22 submission but without access to a reference translation. This means that this model is a multitask model trained on regression and sequence tagging. Both models are ensembled together using a hyperparameter search that weights different features extracted from these two QE models and combines them into a single score.

f200spBLEU (Goyal et al., 2022) computes BLEU over text tokenised with a single language-agnostic SentencePiece subword model. For the f200spBLEU version of spBLEU used in this year's shared task, the SentencePiece tokeniser (Kudo and Richardson, 2018) was trained using data from the FLORES-200 languages.

MS-COMET-22 (Kocmi et al., 2022) is built on top of the COMET (Rei et al., 2020) architecture. It is trained on a set of human judgments several times larger – covering 113 languages and 15 domains. Furthermore, the authors propose filtering out those human judgements with potentially low quality. MS-COMET-22 is a reference-based metric that receives the source, the MT hypothesis and the human reference as input.

Random-sysname is a random baseline. The metric takes the name of the system as the only parameter. It uses a discrete score. Segment-level scores follow a Gaussian distribution around mean value X (in the range 0-9) and a standard deviation of 2. The mean X is calculated from the name of the system as: $X = sha256(sysname)[0]\%10$

The idea behind this baseline is two-fold. Firstly, having a baseline showing how a random metric would perform could help to put scores into context (in particular, pairwise accuracy can create a perception of great performance while 50% is just a toss of a coin). Secondly, it could help to detect errors in metric meta-evaluations.

YiSi-1 (Lo, 2019) measures the semantic similarity between the hypothesis and the reference translation by using cosine similarity scores of multilingual representations at the lexical level. It optionally uses a semantic role labeller to obtain structural similarity. Finally, a weighted F-score based on structural and lexical similarity is used for scoring the hypothesis against the reference translation.

4.2 Metrics Submitted to WMT 2023

We list the descriptions of the metrics submitted to WMT 2023 by the metric developers and refer the reader to the relevant system description papers for further details.

Embed_Llama relies on pretrained Llama 2 embeddings, without any finetuning, to transform sentences into a vector space that establishes connections between geometric and semantic proximities. This metric draws inspiration from Word2vec and utilizes cosine distance for the purpose of estimating similarity or dissimilarity between sentences.

MetricX-23 and MetricX-23-QE are learned reference-based and reference-free (respectively) regression metrics based on the mT5 encoder-decoder language model. They further finetune the mT5-XXL checkpoint on direct assessment data from 2015-2020 and MQM data from 2020 to 2021 as well as synthetic data.

Tokengram_F is an F-score-based evaluation metric that is heavily inspired by chrF++. By replacing word-grams with token-grams obtained from contemporary tokenization algorithms, tokengram_F captures similarities between words sharing the same semantic roots and thus obtains more accurate ratings.

Partokengram_F we did not receive a description of this metric.

XCOMET is a new COMET-base model that is trained to identify errors in sentences along with a final quality score and thus leads to an explainable neural metric. The metric is optimised towards regression and error span detection simultaneously. The same model may be used both with references (XCOMET) and without references (XCOMET-QE). The models are built using XLM-R XL and XXL, thus XCOMET-XL has 3.5B parameters and XCOMET-XXL has 10.7B parameters. The metric is trained in stages where it first sees DAs and then is fine-tuned with MQM. XCOMET-ENSEMBLE is an ensemble between 1 XL and 2 XXL checkpoints that result from these training stages.

XLsim is a supervised reference-based metric that regresses on human scores provided by WMT (2017-2022). Using a cross-lingual language model (XLM-RoBERTa (Conneau et al., 2020)), a supervised model is trained using a Siamese network architecture with CosineSimilarityLoss. **XLSIMQE** is the reference-free variant of this metric.

Cometoid22 is a reference-free metric created using knowledge distillation from reference-based metrics. First, using COMET-22 as a teacher metric, the MT outputs submitted to the WMT News/General Translation task since 2009 are scored. Next, a student metric, called Cometoid22, is trained to mimic the teacher scores without using reference translation. The student metric has the same architecture as COMET-QE, and is initialised with pre-trained weights from the multilingual language model InfoXLM. Three variants were submitted: cometoid22-wmt21,22,23, where the suffix indicates the training data cut-off year.

COMETKiwi-XL and COMETKiwi-XXL use the same COMETKiwi model architecture from WMT 2022 but replace InfoXLM with XLM-R XL and XXL (for COMETKIWI-XL and COMETKIWI-XXL respectively).

KG-BERTScore incorporates a multilingual knowledge graph into BERTScore and generates the final evaluation score by linearly combining the results of KGScore and BERTScore. In contrast to last year, COMET-QE is used to calculate BERTScore.

GEMBA-MQM is an LLM-enabled metric for error quality span marking. It uses three-shot prompting with a GPT-4 model. In contrast to EAPrompt (Lu et al., 2023), it does not require language-specific examples and requires only a single prompt.

5 Results

5.1 Phenomena-level Results

As in Amrhein et al. (2022) we begin by providing a broad overview of metric performance on the different phenomena categories, before conducting more detailed analyses (see Section 5.3). We restrict the overview to the metrics which provide a) segment-level scores and b) scores for all language pairs and directions in ACES. Out of the metrics that participated, 33 fulfil these criteria: 10 baselines, 11 reference-based, and 12 reference-free metrics.

We first compute the Kendall's tau-like correlation scores[3] (Freitag et al., 2021, 2022) for all of the ACES examples. This metric measures the number of times a metric scores the good translation above the incorrect translation (concordant) and equal to or lower than the incorrect translation (discordant):

$$\tau = \frac{concordant - discordant}{concordant + discordant}$$

We then report the average score over all examples belonging to each of the nine top-level accuracy categories in ACES, plus the fluency category *punctuation* (see Table 3). In addition, we compute the ACES-Score, a weighted combination of the top-level categories, which allows us to identify high-level performance trends of the metrics (see Equation 1). Note that the ACES-Score ranges from -29.1 (all phenomena have a correlation of -1) to 29.1 (all phenomena have a correlation of +1).

$$\text{ACES} = sum \left\{ \begin{array}{r} 5 * \tau_{\text{addition}} \\ 5 * \tau_{\text{omission}} \\ 5 * \tau_{\text{mistranslation}} \\ 1 * \tau_{\text{untranslated}} \\ 1 * \tau_{\text{do not translate}} \\ 5 * \tau_{\text{overtranslation}} \\ 5 * \tau_{\text{undertranslation}} \\ 1 * \tau_{\text{real-world knowledge}} \\ 1 * \tau_{\text{wrong language}} \\ 0.1 * \tau_{\text{punctuation}} \end{array} \right\} \quad (1)$$

Overall, the best-performing metrics submitted to this year's shared task, according to the ACES-Score, are COMETKIWI (a reference-free baseline metric), and KG-BERTSCORE (a reference-free metric). BLEU remains one of the worst-performing metrics, with only the random baseline, RANDOM-SYSNAME, achieving a lower ACES-Score. XCOMET-ENSEMBLE is the top ranking among the reference-based metrics. We caution that we developed ACES to investigate strengths and weaknesses of metrics on a phenomena level – hence, we advise the reader not to draw any conclusions based solely on the ACES-Score.

As observed in Amrhein et al. (2022) the performance of the metrics is highly variable, with no clear *winner* in terms of performance across all of the top-level ACES categories. We also observe similar trends in terms of the most challenging categories (*addition*, *undertranslation*, *real-world knowledge*, and *wrong language*). We find that, unlike last year, some metrics perform similarly to or worse than the baseline metrics. In particular, EMBED_LLAMA and GEMBA-MQM which are designed using Large Language Models (LLMs), struggle with this challenge set. This suggests that we need better design strategies in using the rich representations from LLMs for MT evaluation. In general, we find that reference-free metrics perform on par or better than reference-based metrics.

In terms of performance across the top-level categories, we also observe variation in the performance of metrics belonging to the baseline, reference-based, and reference-free groups. The reference-free group exhibits overall stronger performance compared with the other groups, but in particular for the *mistranslation*, *overtranslation*, *undertranslation*, and *real-world knowledge* categories.

[3]Evaluation scripts are available here: https://github.com/EdinburghNLP/ACES

	addition	omission	mistranslation	untranslated	do not translate	overtranslation	undertranslation	real-world knowledge	wrong language	punctuation	ACES-Score
Examples	*999*	*999*	*24457*	*1300*	*100*	*1000*	*1000*	*2948*	*2000*	*1673*	
BERTscore	**0.872**	0.754	0.318	0.771	0.940	-0.186	-0.288	0.030	0.551	**0.844**	9.722
BLEU	0.742	0.427	-0.227	0.353	0.580	-0.838	-0.856	-0.768	0.660	0.704	-2.862
BLEURT-20	0.435	0.812	0.427	0.743	0.860	0.202	0.014	0.388	0.536	0.708	12.048
chrF	0.644	0.784	0.162	**0.781**	**0.960**	-0.696	-0.592	-0.294	0.693	0.773	3.728
COMET-22	0.295	0.822	0.402	0.718	0.820	0.502	0.258	0.382	0.078	0.673	13.458
CometKiwi	0.536	**0.918**	0.614	-0.105	0.520	0.766	0.604	0.577	-0.307	0.765	**17.951**
f200spBLEU	0.666	0.584	-0.082	0.680	0.920	-0.752	-0.794	-0.394	0.657	0.708	0.041
MS-COMET-QE-22	-0.179	0.674	0.440	0.394	0.300	0.524	0.382	0.262	-0.195	0.632	10.027
Random-sysname	-0.117	-0.117	-0.116	-0.083	-0.100	-0.118	-0.152	-0.245	-0.113	-0.074	-3.648
YiSi-1	0.766	0.868	0.354	0.720	0.940	-0.062	-0.076	0.110	0.421	0.763	11.517
eBLEU	0.674	0.682	0.197	0.739	0.880	-0.662	-0.684	-0.042	**0.771**	0.270	3.406
embed_llama	0.211	0.457	0.016	0.503	0.400	-0.170	-0.492	-0.165	0.154	0.476	1.054
MetricX-23	-0.027	0.568	0.578	0.473	0.800	0.790	0.586	0.766	-0.486	0.636	14.091
MetricX-23-b	-0.135	0.622	0.572	0.613	0.860	0.772	0.568	0.749	-0.444	0.532	13.826
MetricX-23-c	-0.015	0.794	0.617	0.611	0.800	0.740	0.526	**0.783**	-0.629	0.527	14.929
partokengram_F	0.087	0.191	-0.034	0.310	0.140	-0.042	-0.028	0.032	0.508	0.171	1.878
tokengram_F	0.698	0.758	0.160	0.779	**0.960**	-0.732	-0.632	-0.273	0.687	0.830	3.492
XCOMET-Ensemble	0.311	0.786	0.663	0.379	0.780	**0.794**	0.612	0.708	-0.423	0.595	17.336
XCOMET-XL	0.169	0.542	0.570	0.222	0.800	0.656	0.464	0.582	-0.367	0.220	13.264
XCOMET-XXL	-0.119	0.413	0.547	0.234	0.600	0.736	0.568	0.508	-0.507	0.509	11.610
XLsim	0.429	0.618	0.153	0.643	0.820	-0.210	-0.290	-0.044	0.392	0.753	5.386
cometoid22-wmt21	-0.339	0.658	0.493	-0.076	0.280	0.670	0.566	0.362	-0.454	0.608	10.409
cometoid22-wmt22	-0.301	0.674	0.493	-0.119	0.280	0.686	0.538	0.340	-0.472	0.599	10.534
cometoid22-wmt23	-0.253	0.702	0.502	-0.046	0.420	0.750	0.590	0.362	-0.319	0.557	11.926
CometKiwi-XL	0.239	0.828	0.624	0.239	0.440	0.762	0.560	0.563	-0.380	0.630	15.988
CometKiwi-XXL	0.361	0.828	0.653	0.414	0.320	0.774	0.560	0.683	-0.537	0.503	16.809
GEMBA-MQM	0.037	0.281	0.153	0.094	0.140	0.466	0.276	0.268	-0.150	0.015	6.419
KG-BERTScore	0.538	0.912	0.585	-0.206	0.700	0.772	0.606	0.594	-0.307	0.654	17.906
MetricX-23-QE	0.045	0.678	0.654	0.379	0.460	0.772	0.612	0.654	-0.702	0.226	14.614
MetricX-23-QE-b	0.027	0.760	0.663	0.489	0.480	0.758	**0.620**	0.647	-0.673	0.256	15.106
MetricX-23-QE-c	-0.115	0.664	**0.721**	0.384	0.340	0.726	0.618	0.753	-0.712	0.375	13.873
XCOMET-QE-Ensemble	0.277	0.754	0.644	0.181	0.720	0.764	0.582	0.626	-0.519	0.449	16.156
XLsimQE	0.205	0.383	0.087	-0.694	0.940	0.454	0.352	0.042	0.307	0.671	8.070
Average	0.232	0.639	0.382	0.349	0.609	0.314	0.187	0.289	-0.069	0.532	10.002

Table 3: Average Kendall's tau-like correlation results for the ACES top-level categories and ACES-Scores (final column). Metrics are grouped into baseline (top), and participating reference-based (middle) and reference-free (bottom) metrics. Note that *Average* is an average over averages. Best results are highlighted in green.

5.2 Mistranslation Results

Next, we drill down to the fine-grained categories of the largest ACES category: *mistranslation*. We present metric performance for the sub-level categories (*discourse*, *hallucination*, and *other*) in Table 4. The *discourse* sub-category includes errors involving the mistranslation of discourse-level phenomena such as pronouns and discourse connectives. *Hallucination* includes errors at the word level that could occur due to hallucination by an MT model, for example, the use of wrong units, dates, times, numbers or named entities, as well as hallucinations at the subword level that result in nonsensical words. The *other* cub-category covers all other categories of mistranslation errors including overly-literal translations (see example in Table 1) and the introduction of ambiguities in the translation output. Again, as in 2022, we find that performance on the different sub-categories is highly variable, with no clear *winner* among the metrics. We also make similar observations to those in Amrhein et al. (2022), that the hallucination phenomena are generally more challenging than discourse-level phenomena; performance on the hallucination sub-category is poor overall.

5.3 Analysis

We repeat the analyses we performed in Amrhein et al. (2022) for the metrics submitted to WMT 2023 to confirm whether our findings from WMT 2022 still hold. We highlight similar observations to those from WMT 2022 and summarise our insights below.

5.3.1 How sensitive are metrics to the source?

Finding: Reference-based metrics tend to ignore the source.

In the ACES *Mistranslation - Ambiguous Translations* category, examples were designed in such a way that given an ambiguous reference the correct translation candidate can only be identified through the source sentence (See an example in Table 9). We leverage this property to present an analysis aimed at discovering how important the source is for different metrics. We exclude from the analysis all metrics that a) do not take the source and b) do not cover all language pairs. This leaves us with 22 metrics: seven reference-based metrics, fourteen reference-free metrics, and the RANDOM-SYSNAME baseline. In Table 5 we present results for the *Ambiguity - Discourse Connectives* (for the ambiguous English discourse connective "since" which can have either causal or temporal meaning), and *Ambiguity - Occupation Names Gender* (male and female) challenge sets.

	disco.	**halluci.**	**other**
Examples	***3698***	***10270***	***10489***
BERTscore	0.563	-0.062	0.361
BLEU	-0.042	-0.418	-0.250
BLEURT-20	0.695	0.141	0.398
chrF	0.406	-0.138	0.160
COMET-22	0.657	0.113	0.383
CometKiwi	0.779	0.465	0.580
f200spBLEU	0.095	-0.190	-0.150
MS-COMET-QE-22	0.631	0.240	0.417
Random-sysname	-0.117	-0.122	-0.111
YiSi-1	0.608	0.017	0.366
eBLEU	0.374	-0.166	0.282
embed_llama	-0.089	-0.140	0.189
MetricX-23	0.757	0.663	0.393
MetricX-23-b	0.749	0.656	0.390
MetricX-23-c	0.694	**0.755**	0.477
partokengram_F	-0.062	-0.101	0.027
tokengram_F	0.396	-0.132	0.157
XCOMET-Ensemble	**0.791**	0.566	0.626
XCOMET-XL	0.706	0.482	0.521
XCOMET-XXL	0.609	0.540	0.504
XLsim	0.217	-0.066	0.236
cometoid22-wmt21	0.782	0.286	0.400
cometoid22-wmt22	0.748	0.290	0.423
cometoid22-wmt23	0.758	0.223	0.478
CometKiwi-XL	0.752	0.501	0.602
CometKiwi-XXL	0.735	0.535	0.661
GEMBA-MQM	0.076	0.291	0.127
KG-BERTScore	0.685	0.466	0.580
MetricX-23-QE	0.728	0.604	0.628
MetricX-23-QE-b	0.694	0.617	0.666
MetricX-23-QE-c	0.747	0.659	**0.739**
XCOMET-QE-Ensemble	0.702	0.558	0.651
XLsimQE	0.053	0.050	0.134
Average	0.511	0.248	0.365

Table 4: Average Kendall's tau-like correlation results for the sub-level categories in mistranslation: **disco**urse-level, **halluci**nation, and **other** errors. Metrics are grouped into baseline (top), and participating reference-based (middle) and reference-free (bottom) metrics. Note that *Average* is an average over averages. Best results are highlighted in green.

In addition, we measure the correlation gain when metrics receive access to disambiguation information via the source – for this we use the *Real-world Knowledge - Commonsense* challenge set i.e. a scenario in which the source contains disambiguation information (See an example in Table 1). In Table 6 we observe that the correlation gain is lower for the majority of the reference-based metric correlation scores compared with the reference-free metric correlation scores, when access to the subordinate clause is provided via the source.

	since		female		male		
	causal	**temp.**	**anti.**	**pro.**	**anti.**	**pro.**	**AVG**
Examples	*106*	*106*	*1000*	*806*	*806*	*1000*	*3824*
Random-sysname	-0.075	-0.019	-0.146	-0.156	-0.109	-0.154	-0.110
COMET-22	-0.868	0.887	-0.254	0.591	-0.467	0.432	0.053
MetricX-23	-1.000	**1.000**	-0.864	-0.062	0.062	0.870	0.001
MetricX-23-b	-1.000	**1.000**	-0.790	0.112	-0.092	0.780	0.002
MetricX-23-c	-0.849	0.849	-0.998	-0.581	**0.576**	**0.996**	-0.001
XCOMET-Ensemble	-0.585	0.981	**0.852**	0.948	0.273	0.922	**0.565**
XCOMET-XL	-0.698	0.906	0.456	**0.960**	-0.330	0.698	0.332
XCOMET-XXL	-0.868	0.925	0.372	0.675	0.541	0.918	0.427
cometoid22-wmt21	-0.698	0.868	0.580	0.950	-0.787	0.022	0.156
cometoid22-wmt22	-0.623	0.868	0.456	0.851	-0.444	0.442	0.258
cometoid22-wmt23	-0.566	0.925	0.342	0.851	0.117	0.844	0.419
CometKiwi	0.075	**1.000**	**0.990**	**0.998**	-0.171	0.440	0.555
CometKiwi-XL	0.075	0.925	0.952	0.990	0.380	0.892	**0.702**
CometKiwi-XXL	**0.132**	0.943	0.932	0.995	0.241	0.796	0.673
GEMBA-MQM	-0.604	0.736	0.722	0.320	-0.762	-0.692	-0.047
KG-BERTScore	0.075	**1.000**	**0.990**	**0.998**	-0.171	0.440	0.555
MS-COMET-QE-22	-0.283	0.811	-0.194	0.323	0.243	0.692	0.265
MetricX-23-QE	-0.472	0.736	0.974	0.995	0.117	0.816	0.528
MetricX-23-QE-b	-0.566	0.868	0.968	0.995	0.722	**0.968**	0.659
MetricX-23-QE-c	-0.302	0.774	0.968	**0.998**	**0.911**	0.866	**0.702**
XCOMET-QE-Ensemble	-0.208	0.925	0.930	0.975	0.546	0.912	0.680
XLsimQE	0.245	-0.113	0.208	0.350	-0.256	-0.170	0.044

Table 5: Results on the challenge sets where the good translation can only be identified through the source sentence. Upper block: reference-based metrics, lower block: reference-free metrics. Best results for each phenomenon and each group of models is marked in bold and green and the average over all can be seen in the last column.

In line with last year's findings, we report that reference-based metrics still lag behind reference-free metrics in terms of their correlation on challenge sets that can only be disambiguated by looking at the source. This indicates that reference-based metrics still rely too much on the reference translation. We conclude that our initial finding from 2022 still holds: that reference-based metrics tend to ignore relevant information in the source. One exception is XCOMET-ENSEMBLE, a reference-based metric that reaches similar correlations and correlation gains as some of the mid-performing reference-free metrics. We suspect that by training the same model to exhibit reference-based and reference-free behaviour, the model learns to utilise the information from the source in addition to the reference, when provided.

5.3.2 How much do metrics rely on surface overlap with the reference?

Finding: Reference-based metrics still rely on reference overlap.

Surface-level metrics are often too reliant on overlap with the reference. We aim to discover whether neural reference-based metrics submitted to the 2023 shared task are able to avoid this problem. Using the *Hallucination - Numbers and Named Entities* challenge set we compared how well reference-based and reference-free metrics[4] on average can identify *number* and *named entity* mismatches. In these challenge sets, we perform both word-level and character-level edits (i.e. substitutions) to simulate the hallucination behaviour. In order to thoroughly understand the behaviour of

[4]Excluding surface-level overlap metrics (BLEU, CHRF, FP200SPBLEU, PARTOKENGRAM_F, TOKENGRAM_F).

	corr. gain
Random-sysname	-0.052
COMET-22	0.042
MetricX-23	0.004
MetricX-23-b	-0.002
MetricX-23-c	0.008
XCOMET-Ensemble	**0.162**
XCOMET-XL	0.110
XCOMET-XXL	0.016
cometoid22-wmt21	0.120
cometoid22-wmt22	0.124
cometoid22-wmt23	0.138
CometKiwi	0.454
CometKiwi-XL	0.148
CometKiwi-XXL	0.108
GEMBA-MQM	**1.107**
KG-BERTScore	0.436
MS-COMET-QE-22	0.198
MetricX-23-QE	0.272
MetricX-23-QE-b	0.296
MetricX-23-QE-c	0.142
XCOMET-QE-Ensemble	0.112
XLsimQE	0.184

Table 6: Results on the *real-world knowledge commonsense challenge set* with reference-based metrics in the upper block and reference-free metrics in the lower block. The numbers are computed as the difference between the correlation with the subordinate clause in the source and the correlation without the subordinate clause in the source. Largest gains are bolded.

metrics under such hallucination errors, we introduced three levels. The first, easiest level follows Freitag et al. (2021) and applies a change to an alternative translation to form an incorrect translation. The second level uses an alternative translation that is lexically very similar to the reference as the good translation and applies a change to the reference to form an incorrect translation. The third, and hardest level, uses an alternative translation that is lexically very different from the reference as the good translation and applies a change to the reference to form an incorrect translation. In this way, our challenge set tests whether number and named entity differences can still be detected as the surface similarity between the two translation candidates decreases and the surface similarity between the incorrect translation and the reference increases. See an example of the different levels below as taken from the dataset paper -

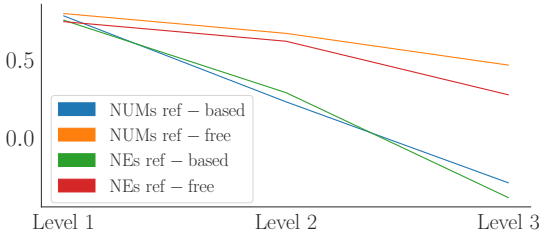


Figure 2: Decrease in correlation for reference-based and reference-free metrics on the named entity and number hallucination challenge sets.

SRC (es): Sin embargo, Michael Jackson, Prince y **Madonna** fueron influencias para el álbum.
REF (en): Michael Jackson, Prince and **Madonna** were, however, influences on the album.

Level-1 ✓: However, Michael Jackson, Prince, and **Madonna** were influences on the album.
Level-1 ✗: However, Michael Jackson, Prince, and **Garza** were influences on the album.

Level-2 ✓: However, Michael Jackson, Prince, and **Madonna** were influences on the album.
Level-2 ✗: Michael Jackson, Prince and **Garza** were, however, influences on the album.

Level-3 ✓: The record was influenced by **Madonna**, Prince, and Michael Jackson though.
Level-3 ✗: Michael Jackson, Prince and **Garza** were, however, influences on the album.

We take the average correlation for all reference-based and reference-free metrics that cover all languages. We then plot the decrease in correlation with increasing surface-level similarity of the incorrect translation to the reference (Figure 2). As in the corresponding analysis of the WMT 2022 metrics, we observe that, on average, reference-based metrics have a much steeper decrease in correlation than the reference-free metrics as the two translation candidates become more and more lexically diverse and the surface overlap between the incorrect translation and the reference increases. This indicates that reference-based metrics may prefer a) an incorrect translation in cases where it is lexically similar to the reference but contains a severe error over b) a good translation that shares little overlap with the reference.

We also observe a clear effect of surface-level overlap between the reference and the hypothesis

	reference-based	reference-free
hallucination	-0.32 ± 0.15	+0.06 ± 0.06
overly-literal	-0.22 ± 0.14	0.00 ± 0.03
untranslated	-0.44 ± 0.11	-0.03 ± 0.06

Table 7: Average correlation difference and standard deviation between the challenge sets with reference-copied good translations and the challenge sets with the synonymous good translations.

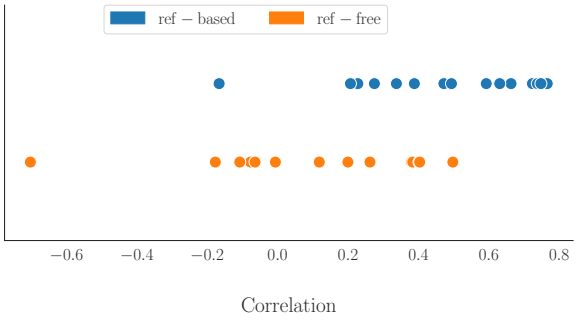


Figure 3: Correlation of reference-based metrics (blue) and reference-free metrics (orange) on the sentence-level untranslated test challenge set.

on three challenge sets for which we have different versions of the good translation, where the error was corrected with: a) the corresponding *correct token* from the reference and b) *a synonym for the correct token* from the reference. In Table 7, we can see a much larger difference in correlation between the challenge sets with reference-copied good translations and the challenge sets with the synonymous good translations, for the reference-based metrics as compared to the reference-free metrics. That is, it is much easier for reference-based metrics to identify mistranslations when the good translation matches a term in the reference compared with when a synonym is used. Furthermore, when the incorrect translation *shares a high degree of lexical overlap with the reference but does not have the same meaning* (as in the *Mistranslation - Lexical Overlap* challenge set based on adversarial paraphrase from PAWS-X (Yang et al., 2019)), the reference-based metrics only reach a correlation of 0.05 ± 0.16 on average. In contrast, the reference-free metrics reach a correlation of 0.27 ± 0.16.

We again conclude that although state-of-the-art reference-based MT evaluation metrics are no longer solely reliant on surface-level overlap, it still has a considerable influence on their predictions.

5.3.3 Do multilingual embeddings help design better metrics?

Finding: Multilingual embeddings can be harmful with poor design.

We are interested in the extent to which the representations in neural MT evaluation metrics, which are trained on multilingual models, are language-dependent. For this analysis, we investigated the effect of alignment of multilingual embeddings (including LLMs) on the evaluation task through the *wrong-language* and *untranslated - full sentences* phenomena for those metrics that provided scores for examples in all language pairs. In the *wrong-language* phenomenon, the incorrect translation contains a high-quality translation of the source in a similar language to the target language while the good translation is the machine translation output of the source sentence in the target language. In the challenge set for *untranslated - full sentences*, the incorrect translation is a copy of the source sentence and the good translation is machine translation output in the target language. Multilingual embeddings learn cross-lingual representations by reducing the language-specific properties during pretraining (Wu and Dredze, 2019). We hypothesised that making representations language agnostic may harm MT evaluation in cases where translations are extremely poor, such that they remain untranslated or hallucinate from a similar language.

In Figure 3 we plot the correlations for all reference-based and reference-free metrics. Overall, we observe that several metrics from 2023 have much better correlation scores than 2022 indicating that newer models have developed strategies to avoid learning language-agnostic representations. In particular, we find that many of the reference-free metrics submitted to the 2023 shared task have improved on the *untranslated - full sentences* category (though a few reference-free metrics from 2022 had performance closer to 1, which is not the case with the 2023 metrics). This is a welcome change as we expect these metrics to perform a more faithful evaluation when many of the words remain untranslated in the hypothesis, especially in the lower resource setting. Whilst some reference-free metrics struggle considerably on this challenge set and almost always prefer the copied source to the real translation, reference-based metrics generally exhibit good correlation i.e. they can identify the copied source quite easily. As reference-based metrics tend to ignore the source, the scores are likely based on the similarity between the reference and the MT output. This is evident from their poor

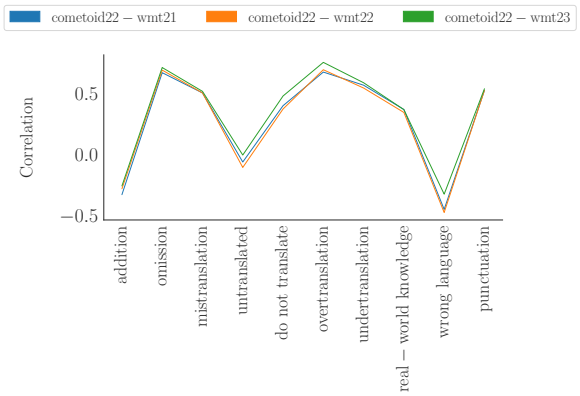


Figure 4: Correlations for different top-level phenomena categories with different models trained on successively more data.

performance on the *wrong - language* category (see Table 3). This suggests that language-agnostic representations present in the multilingual space may harm performance.

5.4 Training data size effects

One submission this year, namely COMETOID22, submitted three different reference-free metric versions, each trained on successively more data. This allows us to investigate the effects of the metric training data size[5] on the performance on ACES. (Note that we cannot draw any conclusions about the training data size of the pretraining models that are used.) In Figure 4 we can see the effect of training data size on the performance on the top-level phenomena categories. COMETOID22-WMT23, the model that has seen the most data, outperforms the other two metrics on almost all top-level categories. The correlation gain is especially pronounced for the *untranslated*, *do not translate* (content in the source is erroneously translated into the target language), *overtranslation* (the target translation contains more specific information than the source) and *wrong language* categories (see Table 9 for examples for each of the phenomena). For clearer insights as to where the performance gain comes from, we would need to analyse the training data in depth. However, it is evident from these results that more training data is beneficial for metric development. In the next section, we look at metric score changes over metric implementation cycles - where likely more than just the training data changed.

[5]Note that for COMETOID22 this is not human judgement labelled data but rather pseudo labelled data where labels come from the reference-based COMET-22 model.

5.5 Changes over one year

We compare the results of metrics submitted by the same teams last year and this year in Table 8.

We report changes in performance in terms of deltas, computed by subtracting the 2022 score from the 2023 score. We do this for the following pairs of metrics: KG-BERTSCORE (2022) and KG-BERTSCORE (2023); COMETKIWI (2022) paired with COMETKIWI-XL (2023) and COMETKIWI-XXL (2023); COMET-22 (2022) paired with XCOMET-ENSEMBLE (2023), XCOMET-XL (2023) and XCOMET-XXL (2023).

We observe that KG-BERTSCORE has improved over its performance of last year. From the description provided by the metric developers, the main difference is that the 2023 version of KG-BERTSCORE metric uses COMET-QE instead of BERTScore (Zhang et al., 2020) to compute the similarity between the source and the hypothesis. Whilst we might therefore attribute the increase in performance to this change, a more systematic comparison of the two metric versions would be required to confirm whether this is the only contributing factor.

The metrics in the COMETKIWI family exhibit: a slight drop in performance (COMETKIWI-XL) and a similar performance to that of last year (COMETKIWI-XXL). The difference can be attributed to changing the underlying encoder, XLM-R XL and XLM-R XXL (Goyal et al., 2021) respectively, and the use of additional fine-tuning data made available this year. We have seen that the addition of more training data helps in Section 5.4. Considering that there is no improvement in the performance, we question if an increase in the underlying model capacity of the encoder alone is useful for obtaining better MT evaluation.

Performance change for the XCOMET family is variable: there is a performance increase for XCOMET-ENSEMBLE (compared to COMET-22), for XCOMET-XL the increase is smaller, and the performance of XCOMET-XXL is degraded. The XCOMET family is designed to provide both a quality score and an error span. Considering that the metric also provides an explanation of the scores without hurting the performance, this is indeed a positive change. Finally, it is worth noting that for *all* metrics in Table 8 a change in performance is observed for almost all ACES categories, for all metrics.

	COMETKiwi		KG-BERTScore	XCOMET		
	-XL	-XXL		-Ensemble	-XL	-XXL
addition	-0.120	-0.004	-0.251	0.595	0.455	0.142
omission	-0.004	-0.002	0.103	0.118	-0.126	-0.254
mistranslation	-0.005	0.013	0.077	0.126	0.038	0.005
untranslated	0.000	0.142	0.266	-0.181	-0.342	-0.362
do not translate	-0.395	-0.553	0.000	0.053	0.079	-0.105
overtranslation	0.027	0.035	0.119	0.073	-0.067	0.017
undertranslation	-0.019	-0.021	0.077	0.014	-0.132	-0.025
real-world knowledge	-0.020	0.100	0.107	0.003	-0.123	-0.198
wrong language	-0.014	-0.173	-0.618	-0.296	-0.232	-0.395
punctuation	-0.037	0.004	0.264	0.206	-0.144	0.006
ACES-Score	-1.04	-0.38	0.40	4.23	0.21	-1.64

Table 8: Comparison of average Kendall's tau-like correlation: delta calculated as 2023 score minus 2022 score.

Whilst it is not possible to draw conclusions or make predictions about the future of metric development based solely on the observations from two consecutive metrics shared tasks, we highlight several high-level changes. Firstly, we note the participation of many more COMET-based metrics in 2023, compared with 2022. This is presumably based on the success of COMET at previous shared tasks and its adoption within the MT community. We find that three metrics from 2022 are now used as baseline metrics namely COMET-22, COMETKIWI, and MS-COMET-QE-22. In contrast to the submissions in 2022, we find some new metrics that use lexical overlap through text matching or embeddings (TOKENGRAM_F, PARTOKENGRAM_F, and EBLEU). However, their performance trend is similar to other surface overlap metrics. This year has also seen submissions based on large language models (EMBED_LLAMA and GEMBA-MQM). As seen in Section 2, their moderate performance indicates the need for more effective approaches. Additionally, we note an overall increase between 2022 and 2023 in the number of metrics submitted to WMT that a) provide segment-level scores and b) provide scores for all language pairs and directions in ACES. There were 37 segment-level metrics at WMT 2022, 24 of which covered the language pairs and directions in ACES, compared with 47 and 33, respectively in 2023. This suggests that the interest in metric development remains high, and could be increasing.

From our analyses in Section 5.3, we also draw similar conclusions to Amrhein et al. (2022) with the exception of reference-free metrics improving at the *Untranslated - Full Sentences* task. Despite the success of LLMs across various tasks (Brown et al., 2020), leveraging them to evaluate translated outputs still requires some improved design strategies. All these observations suggest that evaluating MT outputs is indeed a hard problem (Neubig, 2022). While we do have a good suite of metrics to provide a proxy for evaluation, there are indeed several interesting challenges that need to be tackled before we find an ideal evaluation regime. And even then, we need to continuously monitor this to ensure that we do not optimise towards metric weaknesses that we have not yet discovered.

5.6 Recommendations

We provide the same recommendations as last year:

No metric to rule them all. There is no consistent winning metric across all categories (see Table 3). We recommend developing evaluation methods that combine different design strategies for robust evaluation. We also recommend innovation in the ensemble building as simple strategies like majority voting do not lead to significant improvement (Moghe et al., 2023). We find that some submissions in this year's shared task already contain ensembles (XCOMET-ENSEMBLE, XCOMET-QE-ENSEMBLE) which suggests that our recommendations are in line with the efforts of the community.

The source matters. The trend where reference-based metrics tend to disregard information in the source is also persistent, as seen in Section 5.3. We also observe that reference-free metrics are highly competitive with reference-based metrics as seen in Table 3 and also in Freitag et al. (2022); Zerva et al. (2022), *inter alia*. Furthermore, as references are often not perfect themselves (Freitag et al., 2020), it is ideal to develop evaluation regimens that focus more on the information in the source sentence than

the references.

Surface overlap still prevails. Neural metrics were introduced to overcome surface-level overlap present in the string-based metrics. However, the results in Section 5.3.2 suggest that neural metrics tend to focus more on lexical overlap than semantic content. We thus recommend including paraphrases in the training regime as well as designing loss functions that explicitly discourage surface-level overlap.
Lastly, simple strategies to model language-specific information in the metrics could also improve the robustness of the metrics to language pair attacks.

6 Conclusion

We re-submitted the ACES Challenge Set to WMT2023 to identify the strengths and weaknesses of the metrics submitted to this year's shared task. Overall, we find similar trends to that of last year. While neural metrics tend to be better, different categories of metrics have different strengths, and we do not find one clear winner. With respect to the metrics that were resubmitted with some design changes, we find that these design changes have variable outcomes with a performance drop in some cases. The major challenges of (i) metrics not paying enough attention to the source, (ii) reference-based metrics still relying on surface-level overlap, and (iii) over-reliance on multilingual embeddings still persist. Hence, our recommendations are also similar to that of last year: build ensembles of different design families, encourage development that better utilises information in the source, include diverse training examples to reduce the influence of surface-level overlap, and carefully determine the influence of multilingual embeddings/LLMs on MT evaluation.

Limitations

When comparing the results of the baseline metrics common to the 2022 and 2023 metrics shared tasks, we observed differences in the scores returned for a small subset (2,659; approx 7%) of the ACES examples. A subsequent investigation suggested that differences in the pre-processing steps by the shared task organisers in 2022 and 2023 may have led to the differences; we further conjecture that differences in handling the double quotes present in some of the ACES examples may be one of the main causes. Regardless of the source of the differences, we highlight that care should be taken when pre-processing the ACES dataset prior to benchmarking metric performance, especially when the aim is to draw comparisons with results reported in previous work. However, we note that this issue is not specific to ACES, but may potentially affect any text-based dataset. With the exception of the comparison of results from 2022 and 2023 in Section 5.5, for which we used the subset of 33,817 examples which were unaffected by pre-processing differences, all other results reported in this paper use the full set of 36,476 ACES examples. We also note that ideally, incorrect processing of double quotes by a metric should not lead to a difference in scores especially when dealing with semantic errors.

As we re-submitted the same version of the ACES dataset to WMT 2023, the same biases described in Amrhein et al. (2022) remain: 1) there is greater coverage in terms of phenomena and number of examples for some language pairs (particularly en-de and en-fr), 2) more examples are provided for categories for which examples may be generated automatically, compared to those that required manual construction/filtering, 3) errors present in the datasets used to construct the examples may have propagated through into ACES, 4) the focus of the ACES is on accuracy errors; the inclusion and evaluation of fluency errors remains a direction for future work.

ACES consists of examples that target a range of linguistic phenomena, which are then arranged in a hierarchy of error categories. In order to provide metric profiles over this range of error categories we require segment-level scores. We therefore report only results for those metrics submitted to WMT 2023 that provide segment-level scores; metrics that provide only system-level outputs are excluded. Further, we excluded those metrics that did not provide scores for all of the language pairs in ACES from the results and analyses in this paper.

The 2023 WMT metric shared task evaluated metrics at the paragraph level for English-German. Currently, ACES is not able to capture document-level metric performance. We hope such challenge sets will become available in the near future to be able to track metric improvements beyond the sentence level.

Ethics Statement

As described in Amrhein et al. (2022) some examples within the ACES challenge set exhibit biases.

However, this is necessary in order to expose the limitations of existing metrics. The challenge set is already publicly available.

Acknowledgements

We thank the organisers of the WMT 2023 Metrics task for organising the Challenge Sets shared task, and the shared task participants for scoring our challenge sets with their systems. We thank Nikolay Bogoychev and the anonymous reviewers for their insightful comments and suggestions. This work was supported in part by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh (Moghe), and by the Swiss National Science Foundation (project MUTAMUR; no. 176727) (Amrhein). We also thank Huawei (Moghe) and the RISE Center for Applied AI (Guillou) for their support.

References

Duarte Alves, Ricardo Rei, Ana C Farinha, José G. C. de Souza, and André F. T. Martins. 2022. Robust MT evaluation with sentence-level multilingual augmentation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 469–478, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2022. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Eleftherios Avramidis and Vivien Macketanz. 2022. Linguistically motivated evaluation of machine translation metrics based on a challenge set. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 514–529, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Xiaoyu Chen, Daimeng Wei, Hengchao Shang, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Ting Zhu, Mengli Zhu, Ning Xie, Lizhi Lei, Shimin Tao, Hao Yang, and Ying Qin. 2022. Exploring robustness of machine translation metrics: A study of twenty-two automatic metrics in the WMT22 metric task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 530–540, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. *CoRR*, abs/2105.00572.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán,

and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Liane Guillou and Christian Hardmeier. 2016. PROTEST: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 636–643, Portorož, Slovenia. European Language Resources Association (ELRA).

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics.

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. DEMETR: Diagnosing evaluation metrics for translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9540–9561, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Margaret King and Kirsten Falkedal. 1990. Using test suites in evaluation of machine translation systems. In *COLING 1990 Volume 2: Papers presented to the 13th International Conference on Computational Linguistics*.

Tom Kocmi, Hitokazu Matsushita, and Christian Federmann. 2022. MS-COMET: More and better human judgements improve metric performance. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 541–548, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Yitong Li, Trevor Cohn, and Timothy Baldwin. 2017. BIBI system description: Building with CNNs and breaking with deep reinforcement learning. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 27–32, Copenhagen, Denmark. Association for Computational Linguistics.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Arle Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 0:455–463.

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2023. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt.

Taylor Mahler, Willy Cheung, Micha Elsner, David King, Marie-Catherine de Marneffe, Cory Shain, Symon Stevens-Guille, and Michael White. 2017. Breaking NLP: Using morphosyntax, semantics, pragmatics and world knowledge to fool sentiment analysis systems. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 33–39, Copenhagen, Denmark. Association for Computational Linguistics.

Richard T McCoy and Tal Linzen. 2019. Non-entailed subsequences as a challenge for natural language inference. *Proceedings of the Society for Computation in Linguistics (SCiL)*, pages 358–360.

Nikita Moghe, Tom Sherborne, Mark Steedman, and Alexandra Birch. 2023. Extrinsic evaluation of machine translation metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13060–13078, Toronto, Canada. Association for Computational Linguistics.

Graham Neubig. 2022. Is my nlp model working? the answer is harder than you think.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Maja Popović and Sheila Castilho. 2019. Challenge test sets for MT evaluation. In *Proceedings of Machine Translation Summit XVII: Tutorial Abstracts*, Dublin, Ireland. European Association for Machine Translation.

Abhilasha Ravichander, Siddharth Dalmia, Maria Ryskina, Florian Metze, Eduard Hovy, and Alan W Black. 2021. NoiseQA: Challenge set evaluation for user-centric question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2976–2992, Online. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Guido Rocchietti, Flavia Achena, Giuseppe Marziano, Sara Salaris, and Alessandro Lenci. 2021. Fancy: A diagnostic data-set for nli models. In *Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it)*.

Thibault Sellam, Amy Pu, Hyung Won Chung, Sebastian Gehrmann, Qijun Tan, Markus Freitag, Dipanjan Das, and Ankur Parikh. 2020. Learning to evaluate translation beyond English: BLEURT submissions to the WMT metrics 2020 shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 921–927, Online. Association for Computational Linguistics.

Ieva Staliūnaitė and Ben Bonfil. 2017. Breaking sentiment analysis of movie reviews. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 61–64, Copenhagen, Denmark. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Examples from ACES

We shall now list one example from every top-level category in Table 9. We reuse most of the examples mentioned in the original paper under the respective categories.

	Addition *target includes content not present in the source*
SRC (de):	In den letzten 20 Jahren ist die Auswahl in Uptown Charlotte exponentiell gewachsen.
REF (en):	In the past 20 years, the amount in Uptown Charlotte has grown exponentially.
✓:	Over the past 20 years, the selection in Uptown Charlotte has grown exponentially.
✗:	Over the past 20 years, the selection of **child-friendly options** in Uptown Charlotte has grown exponentially.
	Omission *errors where content is missing from the translation that is present in the source*
SRC (fr):	Une tornade est un tourbillon d'air à basse-pression en forme de colonne, l'air alentour est aspiré vers l'intérieur et le haut.
REF (en):	A tornado is a **spinning column** of very low-pressure air, which sucks the surrounding air inward and upward.
✓:	A tornado is a **column-shaped** low-pressure air turbine, the air around it is sucked inside and up.
✗:	A tornado is a low-pressure air turbine, the air around it is sucked inside and up.
	Untranslated - Word Level *errors occurring when a text segment that was intended for translation is left untranslated in the target content*
SRC (fr):	À l'origine, l'émission mettait en scène des **comédiens de doublage** amateurs, originaires de l'est du Texas.
REF (de):	Die Sendung hatte ursprünglich lokale Amateur**synchronsprecher** aus Ost-Texas.
✓ (copy):	Ursprünglich spielte die Show mit Amateur**synchronsprechern** aus dem Osten von Texas.
✓ (syn.):	Ursprünglich spielte die Show mit Amateur-**Synchron-Schauspielern** aus dem Osten von Texas.
✗:	Ursprünglich spielte die Show mit Amateur-**Doubling-Schauspielern** aus dem Osten von Texas.
	Mistranslation - Ambiguous Translation *an unambiguous source text is translated ambiguously*
SRC (de):	Der Manager feuerte **die** Bäcker**in**.
REF (en):	The manager fired the baker.
✓:	The manager fired the **female** baker.
✗:	The manager fired the **male** baker.
	Do Not Translate *content in the source that should be copied to the output in the source language, but was mistakenly translated into the target language.*
SRC (en):	Dance was one of the inspirations for the exodus - song **"The Toxic Waltz"**, from their 1989 album "Fabulous Disaster".
REF (de):	Dance war eine der Inspirationen für das Exodus-Lied **„The Toxic Waltz"** von ihrem 1989er Album „Fabulous Disaster".
✓:	Der Tanz war eine der Inspirationen für den Exodus-Song **„The Toxic Waltz"**, von ihrem 1989er Album „Fabulous Disaster".
✗:	Der Tanz war eine der Inspirationen für den Exodus-Song **„Der Toxische Walzer"**, von ihrem 1989er Album „Fabulous Disaster".
	Undertranslation *erroneous translation has a meaning that is more generic than the source*
SRC (de):	Bob und Ted waren Brüder. Ted ist der **Sohn** von John.
REF (en):	Bob and Ted were brothers. Ted is John's **son**.
✓:	Bob and Ted were brothers, and Ted is John's **son**.
✗:	Bob and Ted were brothers. Ted is John's **male offspring**.
	Overtranslation *erroneous translation has a meaning that is more specific than the source*
SRC (ja):	その 40 分の映画はアノーがアラン・ゴダードと協力して脚本を書いた。
REF (en):	The 40-minute **film** was written by Annaud with Alain Godard.
✓:	The 40-minute **film** was written by Annaud along with Alain Godard.
✗:	he 40-minute **cinema verite** was written by Annaud with Alain Godard.
	Real-world Knowledge - Textual Entailment *meaning of the source/reference is entailed by the "good" translation*
SRC (de):	Ein Mann **wurde ermordet**.
REF (en):	A man **was murdered**.
✓:	A man **died**.
✗:	A man **was attacked**.
	Wrong Language *incorrect translation is a perfect translation in a related language*
SRC (en):	Cell comes from the Latin word cella which means small room.
REF (es):	El término célula deriva de la palabra latina cella, que quiere decir «cuarto pequeño».
✓ (es):	La célula viene de la palabra latina cella que significa habitación pequeña.
✗ (ca):	Cèl·lula ve de la paraula llatina cella, que vol dir habitació petita.

Table 9: Examples from each top-level accuracy error category in ACES. An example consists of a source sentence (SRC), reference (REF), good (✓) and incorrect (✗) translations, language pair, and a phenomenon label. We also provide a description of the relevant phenomenon. en: English, de: German, fr: French, ja: Japanese, es: Spanish, ca: Catalan

Challenging the state-of-the-art Machine Translation Metrics from a Linguistic Perspective

Eleftherios Avramidis, Shushana Manakhimova, Vivien Macketanz and Sebastian Möller
German Research Center for Artificial Intelligence (DFKI),
Speech and Language Technology, Berlin, Germany
firstname.lastname@dfki.de

Abstract

We employ a linguistically motivated challenge set in order to evaluate the state-of-the-art machine translation metrics submitted to the Metrics Shared Task of the 8th Conference for Machine Translation. The challenge set includes about 21,000 items extracted from 155 machine translation systems for three language directions (German⇔English, English→Russian), covering more than 100 linguistically-motivated phenomena organized in 14 categories. The metrics that have the best performance with regard to our linguistically motivated analysis are the COMETOID22-WMT23 (a trained metric based on distillation) for German-English and METRICX-23-C (based on a fine-tuned mT5 encoder-decoder language model) for English-German and English-Russian. Some of the most difficult phenomena are *passive voice* for German-English, *named entities, terminology* and *measurement units* for English-German and *focus particles, adverbial clause* and *stripping* for English-Russian.

1 Introduction

Most NLP evaluation has relied for years on testing the system performance on randomly picked test sets and producing a single generic score. Yet, machine learned systems learn to make abstractions and due to these, phenomena who are on the long tail of the training and test data may be overlooked hidden behind a very high generic score. Additionally, generic scores are often helpful to show relative improvement and reflect overall quality, but cannot explain the performance in a comprehensive way.

For example, old-style machine translation (MT) metrics measuring lexical overlap would equally penalize the omission of an article and the omission of the particle forming the negation in a sentence, although negation is more crucial for its meaning. While the evaluation of so obvious errors has been addressed by the trained MT metrics, their evaluation relies on correlations with human judgments on randomly picked test-sets. In this case, a single correlation score may not be able to explain the strengths and weaknesses of the metrics with regard to the functioning of language.

Motivated by these considerations, we employ a multifold test set with linguistically-motivated challenges that will allow us to understand the metric performance from a linguistic perspective. These challenges are organized in smaller sets, one set per phenomenon, whereas the phenomena are organized in broader categories. By measuring the ability of the metrics to detect the errors in these challenge sets, we can get scores that indicate different aspects of linguistic performance.

This paper describes the application of such a challenge set on the evaluation of the MT metrics submitted at the relevant shared task of the 8th Conference of Machine Translation (Freitag et al., 2023). The rest of the paper is structured as following: Section 2 describes related work, and section 3 describes the way the challenges were selected. In Section 4 the results are presented and described, first from the perspective of metric comparison and then focusing on the performance for particular linguistically-motivated categories and phenomena per language direction. Some conclusions are given in Section 5.

2 Related work

There has been a growing interest for more fine-grained evaluation of Natural Language Processing (NLP) tools, as shown by the increasing number of publications many of whom have received distinctions (Ribeiro et al., 2020; Avelino et al., 2022; Campolungo et al., 2022). Concerning machine translation (MT), initial efforts were made in the 1990s with the introduction of test suites (King and Falkedal, 1990), and these efforts have been revitalized in light of recent advancements in the

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 713–729
December 6–7, 2023.

field (Guillou and Hardmeier, 2016). To the best of our knowledge, the first endeavours related to the use of challenge sets in a meta-level in order to evaluate MT metrics were applied to Quality Estimation metrics (Avramidis et al., 2018), based on the first version of our linguistically-motivated test suite (Macketanz et al., 2018). The analysis was broadened to cover a broader range of MT metrics, including reference-based ones, as appeared in the Findings paper of the Metrics shared task of the 6th Conference on Machine Translation (Freitag et al., 2021), which was based on a later version of our test suite on German-English (Avramidis et al., 2019, 2020; Macketanz et al., 2021, 2022a), a resource also employed in this paper.

With the occasion of the first challenge set subtask for the metrics shared task of the 7th Conference on Machine Translation (Freitag et al., 2022), a few more challenge sets emerged. ACES (Amrhein et al., 2022) for example, focuses on 68 accuracy errors. Similarly, Alves et al. (2022) evaluate the robustness of MT metrics by generating translations with critical errors. In a more linguistic direction, Chen et al. (2022) examine the capability of the metrics to correlate synonyms in different areas and to discern catastrophic errors at both word- and sentence-levels.

Our submission at that sub-task (Avramidis and Macketanz, 2022) augmented the preliminary analysis appearing at Freitag et al. (2021) by adding the language direction of English-German and presenting a more fine-grained analysis, not only in the category level but also on the phenomenon level. This year's submission, explained on our paper, includes that same challenge set as last year, whereas English-Russian has been added as an additional language direction.

3 Method

3.1 Test suite for MT systems

Here, we are going to explain how we created the pool of MT sentences that were used for the challenge set. The selection was based on a linguistically-motivated test suite (Macketanz et al., 2022a)[1]. The test suite contains a set of source sentences focusing on particular phenomena, each of them accompanied by some rules or regular expressions that can detect which translations would be accepted for these source sentences. This allows a semi-automatic evaluation when new translations are provided, whereas a human annotator resolves cases not covered by the rules.

For this experiment, we employed the test suite on three language directions: German-English (Avramidis et al., 2020), English-German (Macketanz et al., 2021) and English-Russian (Macketanz et al., 2022b). The German-English side consists of 5,539 German test sentences covering 107 linguistically motivated phenomena, the English-German side consists of 4,782 English test sentences covering 126 phenomena, and the English-Russian side consists of 1,225 English test sentences covering 64 phenomena. All language directions are organized in 14 categories, which nevertheless differ among the directions.

The above described test suite has been used to evaluate the outputs of 116 German-English, 29 English-German systems and 10 English-Russian systems submitted at the translation task of the Conference of Machine Translation (WMT). German-English outputs were collected from systems submitted in the years 2018-2021, English-German outputs in the years 2020-2021 and English-Russian in 2022.

3.2 Challenge set for MT metrics

The sentences selected with the help of the test suite are consequently used to create the challenge set. The source sentences and the system outputs have to be organized in contrastive pairs of correct/incorrect translations and a reference. In order to achieve this, for every source sentence from the test suite selection we create a challenge item including:

- one correct translation to be used as a reference translation,
- another correct translation to be used as the first translation candidate
- one incorrect translation to be used as the contrastive translation candidate

The two candidate translations and the reference consist one challenge item. Since source and translations were collected as a result of testing for a particular phenomenon, the same phenomenon will be what the challenge item will test.

Given that we may have many correct and wrong translations for the same source, the reference and the translations of the challenge items result from random combinations of correct and wrong translations from the collected WMT outputs. Therefore,

[1] https://github.com/DFKI-NLP/mt-testsuite

the same source sentence may appear many times.

As a result, we get a challenge set with 10,402 items for German-English, 8,945 items for English-German and 1,727 items for English-Russian.

3.3 Evaluation of metrics

For each challenge item, the two machine translation (MT) outputs, are provided to the metrics as separate MT hypotheses. Which output is correct, and which is incorrect, is hidden from the metrics. These hypotheses are then evaluated against the previously mentioned reference and/or the source. An item is deemed correctly scored when the metric assigns a higher score to the correct MT output compared to the incorrect one. Following this, the statistics below are computed:

i) **Accuracy per Phenomenon**: the ratio of all correctly-scored challenge items per phenomenon to the total number of challenge items for that particular phenomenon.

ii) **Accuracy per Category**: the ratio of all correctly-scored challenge items per category to the total number of challenge items for that category, after consolidating the underlying phenomena of that category into a single set.

Significance tests are performed to compare the highest metric accuracy for each phenomenon with all other metric accuracies for the same phenomenon. This is a one-tailed Z-test, conducted with a significance level of $\alpha = 0.95$. Metrics with accuracies that are not significantly worse than the highest accuracy are considered to share the top position for that phenomenon. A similar approach is used to identify the best accuracies per category, after aggregating the challenge items from the underlying phenomena within each category.

Metric categories We conduct this significance testing in two stages: first, for all metrics involved in the shared task, and then separately for each of the three metric categories (baseline, Quality Estimation (QE) as a metric, reference-based). Systems that are significantly superior per phenomenon across all metrics are highlighted with a gray background, while those that are significantly superior per metric category are denoted in boldface.

Averaging Lastly, we provide three types of averaging scores:

i) **Micro-average:** This approach considers all items equally, aggregating all test items to compute the average percentages.

ii) **Category macro-average:** Here, all categories are treated equally, with the percentages being computed independently for each category and then averaged.

iii) **Phenomenon macro-average:** This average treats all phenomena equally, with the percentages being computed independently for each phenomenon and then averaged.

4 Results

The results are displayed in detail in Tables 1, 2 and 3 for the category level and in Tables 4, 5 and 6 for the phenomenon level, for the three language directions respectively.

4.1 Metric performance analysis

Here we are observing the statistics with a focus on comparing the performance of various metrics on the challenge set.

German-English The accuracies of the metrics, as measured for several categories in German-English, can be seen in Table 1. The best performing metric for German-English is COMETOID22-WMT23 (Gowda et al., 2023), which, wins significantly based on both the micro-average (83%) and the macro-average (87%). This metric is a distilled QE model that has been trained on COMET (Rei et al., 2020) scores of WMT outputs, including the ones of WMT23. For this reason, we include it into the reference-aware metrics. We notice that its performance among the other metrics is impressive. It is the first metric in 6 categories and among them the only one who wins at *Verb tense/aspect/mood* and *function words*, achieving 93% and 91% accuracy respectively.

Another two reference-based baseline metrics, COMET and PRISMREF (Thompson and Post, 2020a,b) share the first position when the category macro-average is considered (82%). None of the other reference-aware metrics submitted this year managed to compete with the metrics with the highest accuracy mentioned above.

The lowest performing metric is the reference-less random baseline RANDOM-SYSNAME, provided by the organizers (44%), followed by XL-SIMQE (55-58%; Mukherjee and Shrivastava,

2023) and MATESE (57-58%; Perrella et al., 2022).

When considering the metric performance with regard to particular categories, one can see, again this year, that different metrics win in different combinations of categories. Here, only COMETOID22-WMT23 as mentioned above, wins 6 metrics, followed by PRISMREF and METRIC-23-C, which win 4 categories. 17 metrics do not win any category, ranging in accuracies around 75%.

English-German The accuracies of the metrics, as measured for several categories in English-German, can be seen in Table 2. The best performing metric in English-German is METRICX-23-C, which is in the first significance cluster based on both the micro-average (81%) and the category macro-average (84%). This metric uses the mT5 encoder-decoder language model, which is fine-tuned using direct assessment data, MQM (Lommel et al., 2014) data and synthetic data. The categories to which its success may be mostly attributed are the *multi-word expressions* (MWE; with 92%) and the non-verbal agreement (95%).

Another three metrics share the first position, when the micro average is considered, namely the QE version of the latter, *MetricX-23-QE-c* and also *mbr-metricx-qe* (Naskar et al., 2023) and XCOMET-Ensemble. It is impressive that QE methods manage to reach high accuracy without access to reference content.

When looking at the worse-performing metrics, MATESE here performs worse than the baseline (36-38%), followed by PARTOKENGRAM_F (55-56%; Dreano et al., 2023b).

In English-German it is even harder to say which metrics perform well in multiple categories, as only one of them, XCOMET-QE-ENSEMBLE, achieves the highest performance in 3 categories (*function words, non-verbal agreement and subordination*). The rest of the metrics show a good performance in 2 categories or fewer.

English-Russian The accuracies of the metrics, as measured for several categories in English-German, can be seen in Table 3. For this language pair, variants of the MetricX achieve significantly higher accuracies than all the other metrics. In particular, METRICX-23-C achieves the highest accuracy based on both micro-average and category macro-average, whereas METRICX-23-B and METRICX-23-QE-C achieve a slightly lower macro-average, which is nevertheless not significantly worse than the one of the former. MATESE is again by far the lowest performing metric (32/34%), lower than the random baseline. We may assume that this metric has not been optimized for this language direction.

4.2 Linguistically motivated analysis

In this section, we are focusing on the results for particular phenomena or categories.

4.2.1 German-English

Category-level The overall average accuracy of all metrics with regard to the linguistically motivated categories is at 76% for German-English, which is two percentage points lower than last year's average. It is still a fact, that the metrics in average fail to predict properly the scores for one out of four challenge items that we provided. Luckily, there has been noticeable accuracy for some categories, for example METRICX reference-based variants achieved an accuracy of 96% for *false friends*, whereas *negation* errors have been scored correctly with a 98.5

The worse performing category is *Verb valency*, where the best metrics achieved only 66% accuracy, and the rest of the metrics averaged to a mere 56%. In this category one can observe the lowest accuracy, given by an LLM-based metric, EMBED_LLAMA (Dreano et al., 2023a) with 41%.

Phenomenon-level The best accuracy for this language pair (Table 4) is achieved this year at several variations of verb tenses, i.e. *Transitive - future II, Modal negated - present, Reflexive - preterite subjunctive II* and *Intransitive - pluperfect* which get more than 85% in average.

The lowest accuracy of all metrics in average is given for *passive voice*, where the highest accuracy achieved by several metrics is only 54.5%. Errors related to *commas, domain-specific terms* and *locations* have also been scored with a less than 65% accuracy.

4.2.2 English-German

Category-level The overall average accuracy of all metrics with regard to the linguistically motivated categories is at 71-73% for English-German. The category where all metrics perform better in average is *negation* (83%), where 11 of the metrics achieve more than 90% accuracy. Negation is closely followed by *function words Non-verbal agreement* (80%).

The worse performing category in average is *named entities and terminology* (58,8%), where most metrics' accuracies are close to 50%, except for BLEURT (Yan et al., 2023) (80.3%). The rest of the categories lie in rather mediocre accuracies, between 58.8% and 80%.

Phenomenon-level The English-German phenomena, where metrics perform best in average (Table 5) are the *transitive conditional II simple, gerunds, contact clause* and the *intransitive present perfect simple*, achieving more than 85% of accuracy. The phenomena which incur the lowest average accuracies are the *transitive present progressive, measuring units*, *modals* and *intransitive - future II progressive* with less than 50% accuracy. The former and the latter were observed as the most difficult phenomena to score also last year.

4.3 English-Russian

This analysis for English-Russian occurs for the first time this year, based on the MT outputs collected at last year's shared task. For this purpose the test instances are much fewer than the other language pairs and therefore the numbers are not very conclusive. Therefore, categories and phenomena that have only a handful of samples will not be included in our analysis, although they appear in the tables.

Category-level Here, the average accuracy over all metrics is much lower than the other language directions, reaching only 66%, only 20% above the random baseline. The best performing category is *ambiguity* (86,3%), more than 13% better than the following categories. The worst performing categories are *function words* and *punctuation*, with less than 55%. The rest of the categories range in accuracies between 53 and 73%.

Phenomenon-level The good performance of the *ambiguity* category is also confirmed in the table on the phenomenon level (Table 6), as in Russian this is the only phenomenon of this category, as opposed to other language pairs where we also have examples of structural ambiguity. The most difficult phenomena to score appear to be the *focus particles, adverbial clause* and *stripping* with less than 50% average accuracy, in many cases lower than the random baseline.

5 Conclusion

In this paper we analysed the performance of several state-of-the-art metrics with regard to particular linguistically-motivated phenomena for three language pairs, German-English, English-German and for the first time, English Russian. The analysis gave a multitude of observations, regarding both the performance of the metrics and the corresponding linguistic observations.

The metrics demonstrating the best performance in average are COMETOID22-WMT23 for the German-English language pair, and METRICX-23-C for both the English-German and English-Russian language pairs. Quality estimation methods have impressively good performance in several phenomena. Some metrics that report usage of LLMs (EMBED_LLAMA) have not scored very high in overall, indicating that more work is required in this direction.

Among the various linguistic phenomena, we could identify some of the particularly challenging ones. In German-English, metrics have difficulties scoring the *passive voice* properly. In English-German *named entities and terminology* as well as specific *measurement units* pose the most difficulties. In English-Russian translation, translations with *focus particles*, *adverbial clause*, and *stripping* phenomena emerge as particularly complex challenges.

Acknowledgements

This research was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft; DFG) through the project TextQ and by the German Federal Ministry of Education (BMBF) through the project SocialWear (grant num. d01IW20002). We would like to thank Hans Uszkoreit, Aljoscha Burchardt, Ursula Strohriegel, Renlong Ai, He Wang, Ekaterina Lapshinova-Koltunski and Sergei Bagdasarov for their prior contributions for the creation of the test suite.

References

Duarte Alves, Ricardo Rei, Ana C Farinha, José G. C. de Souza, and André F. T. Martins. 2022. Robust MT evaluation with sentence-level multilingual augmentation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 469–478, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2022. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Mariana Avelino, Vivien Macketanz, Eleftherios Avramidis, and Sebastian Möller. 2022. A Test Suite for the Evaluation of Portuguese-English Machine Translation. In *Computational Processing of the Portuguese Language*, pages 15–25, Cham. Springer International Publishing.

Eleftherios Avramidis and Vivien Macketanz. 2022. Linguistically motivated evaluation of machine translation metrics based on a challenge set. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 514–529, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Eleftherios Avramidis, Vivien Macketanz, Arle Lommel, and Hans Uszkoreit. 2018. Fine-grained evaluation of quality estimation for machine translation based on a linguistically motivated test suite. In *Proceedings of the AMTA 2018 Workshop on Translation Quality Estimation and Automatic Post-Editing*, pages 243–248, Boston, MA. Association for Machine Translation in the Americas.

Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, Aljoscha Burchardt, and Sebastian Möller. 2020. Fine-grained linguistic evaluation for state-of-the-art machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 346–356, Online. Association for Computational Linguistics.

Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, and Hans Uszkoreit. 2019. Linguistic evaluation of German-English machine translation using a test suite. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 445–454, Florence, Italy. Association for Computational Linguistics.

Niccolò Campolungo, Federico Martelli, Francesco Saina, and Roberto Navigli. 2022. DiBiMT: A novel benchmark for measuring Word Sense Disambiguation biases in Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4331–4352, Dublin, Ireland. Association for Computational Linguistics.

Xiaoyu Chen, Daimeng Wei, Hengchao Shang, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Ting Zhu, Mengli Zhu, Ning Xie, Lizhi Lei, Shimin Tao, Hao Yang, and Ying Qin. 2022. Exploring robustness of machine translation metrics: A study of twenty-two automatic metrics in the WMT22 metric task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 530–540, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Sören Dreano, Derek Molloy, and Noel Murphy. 2023a. Embed_Llama: using LLM embeddings for the Metrics Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Sören Dreano, Derek Molloy, and Noel Murphy. 2023b. Tokengram_F, a fast and accurate token-based chrF++ derivative. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thomson, Tom Kocmi, Frédéric Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Thamme Gowda, Tom Kocmi, and Marcin Junczys-Dowmunt. 2023. Cometoid: Distilling Strong Reference-based Machine Translation Metrics into Even Stronger Quality Estimation Metrics. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Liane Guillou and Christian Hardmeier. 2016. PROTEST: A Test Suite for Evaluating Pronouns in Machine Translation. *Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Margaret King and Kirsten Falkedal. 1990. Using test suites in evaluation of machine translation systems.

In *Proceedings of the 13th conference on Computational Linguistics*, volume 2, pages 211–216, Morristown, NJ, USA. Association for Computational Linguistics.

Arle Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. 2014. Using a new analytic measure for the annotation and analysis of MT errors on real data. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pages 165–172, Dubrovnik, Croatia. European Association for Machine Translation.

Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, and Hans Uszkoreit. 2018. Fine-grained evaluation of German-English machine translation based on a test suite. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 578–587, Belgium, Brussels. Association for Computational Linguistics.

Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, He Wang, Renlong Ai, Shushen Manakhimova, Ursula Strohriegel, Sebastian Möller, and Hans Uszkoreit. 2022a. A linguistically motivated test suite to semi-automatically evaluate German–English machine translation output. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 936–947, Marseille, France. European Language Resources Association.

Vivien Macketanz, Eleftherios Avramidis, Shushen Manakhimova, and Sebastian Möller. 2021. Linguistic evaluation for the 2021 state-of-the-art machine translation systems for German to English and English to German. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1059–1073, Online. Association for Computational Linguistics.

Vivien Macketanz, Shushen Manakhimova, Eleftherios Avramidis, Ekaterina Lapshinova-koltunski, Sergei Bagdasarov, and Sebastian Möller. 2022b. Linguistically motivated evaluation of the 2022 state-of-the-art machine translation systems for three language directions. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 432–449, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ananya Mukherjee and Manish Shrivastava. 2023. MEE4 and XLsim : IIIT HYD's Submissions' for WMT23 Metrics Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Subhajit Naskar, Daniel Deutsch, and Markus Freitag. 2023. Quality Estimation using Minimum Bayes Risk. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Niccolò Campolungo, and Roberto Navigli. 2022. MaTESe: Machine translation evaluation as a sequence tagging problem. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 569–577, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020a. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020b. Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the Fifth Conference on Machine Translation*, pages 561–570, Online. Association for Computational Linguistics.

Yiming Yan, Tao Wang, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Mingxuan Wang. 2023. BLEURT has universal translations: An analysis of automatic metrics by minimum risk training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5428–5443, Toronto, Canada. Association for Computational Linguistics.

		baselines								QE as a metric													ref. based metrics																
ling. category	\#	BERTscore	BLEU	BLEURT-20	COMET	YiSi-1	chrF	prismRef	spBLEU	CometKiwi-XL	CometKiwi-XXL	CometKiwi	GEMBA-MQM	MS-COMET-QE	MetricX-23-QE-b	MetricX-23-QE-c	MetricX-23-QE	Random-sysname	XCOMET-QE-Ensemble	XLsimQE	cometoid22-wmt21	cometoid22-wmt22	prismSrc2	KG-BERT	MEE4	MEE4_stsb_xlm	MaTESe	MetricX-23-b	MetricX-23-c	MetricX-23	XCOMET-Ensemble	XCOMET-XL	XCOMET-XXL	XLsim	cometoid22-wmt23	eBLEU	embed_llama	partokengram_F	tokengram_F	avg
Ambiguity	298	87	71	89	**90**	**90**	81	**91**	79	78	**90**	85	77	82	78	86	80	46	87	61	86	85	66	85	83	88	74	87	90	90	88	89	**92**	84	**91**	74	56	48	80	81
Composition	252	**90**	66	88	87	88	74	86	71	83	87	80	81	75	85	**90**	81	44	86	40	81	81	70	80	85	86	69	85	**90**	85	87	84	83	79	87	74	65	61	75	79
Coordination & ellipsis	316	81	74	81	78	**83**	78	**82**	74	**83**	75	**84**	69	67	78	77	78	45	79	59	75	73	79	**84**	81	77	49	81	**82**	81	81	79	70	77	81	73	68	58	78	75
False friends	90	90	64	**91**	**92**	88	76	84	69	82	82	82	77	72	**89**	**90**	88	47	80	51	87	83	86	82	87	80	73	**96**	**96**	**96**	84	84	74	74	93	72	68	51	72	80
Function word	586	82	73	**83**	**83**	82	73	**84**	73	84	**87**	76	77	68	80	86	80	47	86	60	84	**87**	81	76	79	76	57	87	83	86	87	83	72	80	**91**	79	66	55	75	78
LDD & interrogatives	1014	84	74	85	**87**	84	76	**86**	74	80	81	**84**	72	81	**83**	**83**	**82**	45	80	61	81	79	72	84	81	76	53	82	83	82	84	82	72	77	**87**	76	68	54	77	77
MWE	610	82	72	84	84	**85**	76	**86**	73	74	83	77	73	59	80	**87**	80	45	83	53	72	73	65	76	80	81	67	84	86	**87**	84	84	80	75	85	75	66	54	76	76
Named entity & termin.	861	73	64	69	74	**76**	69	73	72	65	67	66	61	62	65	**70**	66	44	67	59	66	67	**69**	66	72	70	58	72	74	72	72	**75**	66	69	71	64	57	52	70	67
Negation	76	79	78	91	**93**	84	82	84	76	91	91	93	75	84	91	**95**	86	50	93	55	91	89	92	93	86	84	51	91	**95**	92	**95**	91	79	76	91	75	59	45	82	82
Non-verbal agreement	419	78	78	**85**	**84**	78	78	79	79	78	**80**	77	67	78	78	**81**	79	43	**80**	64	77	79	74	76	78	75	59	81	**84**	81	81	77	72	70	81	76	60	74	77	76
Punctuation	293	75	74	69	69	70	65	**80**	76	62	62	63	59	59	59	61	60	39	65	67	66	66	**86**	63	**76**	67	48	62	55	62	66	66	44	69	67	65	66	67	67	65
Subordination	679	75	70	76	**78**	75	70	**78**	69	**79**	**78**	**78**	64	65	77	77	**78**	43	**79**	53	75	75	68	**78**	77	75	50	77	78	77	**79**	75	64	70	**79**	73	62	71	71	72
Verb tense/aspect/mood	4697	88	69	85	88	**89**	76	**90**	71	83	84	85	84	81	85	**87**	81	43	82	60	74	75	80	85	84	84	56	82	87	83	85	83	78	78	**93**	76	62	77	77	79
Verb valency	211	**64**	54	63	**64**	64	55	**65**	52	**59**	54	**59**	45	51	**60**	58	55	39	57	33	57	**59**	48	59	62	56	44	62	63	61	62	61	53	61	**66**	55	41	56	56	56
macro avg.	10402	81	70	**81**	**82**	**81**	73	**82**	72	77	79	78	70	70	78	**81**	77	44	79	55	76	77	74	78	79	77	58	81	82	81	81	79	71	74	**83**	72	62	59	74	74
micro avg.	10402	83	70	82	**84**	**84**	75	**85**	72	79	80	80	75	74	80	**83**	78	44	80	58	75	76	75	80	81	79	57	81	83	81	82	81	74	76	**87**	74	63	67	75	76

Table 1: Accuracy of the metrics (%) with regard to the 14 linguistically motivated categories for German-English. The significantly best systems per phenomenon over all metrics are indicated with a gray background, whereas the significantly best systems per metrics category are indicated with boldface.

		baselines								QE as a metric																					ref. based metrics																				
ling. category	#	BERTscore	BLEU	BLEURT-20	COMET	YiSi-1	chrF	prismRef	spBLEU	Calibri-COMET22-QE	CometKiwi-XL	CometKiwi-XXL	CometKiwi	GEMBA-MQM	KG-BERT	MS-COMET-QE	MetricX-23-QE-b	MetricX-23-QE-c	MetricX-23-QE	Random-sysname	XCOMET-QE-Ensemble	XLsimQE	cometoid22-wmt21	cometoid22-wmt22	cometoid22-wmt23	mbr-metricx-qe0p2p1-qe	mbr-metricx-qexv1p-qe	prismSrc2	Calibri-COMET22	MEE4	MEE4_stsb_xlm	MaTESe	MetricX-23-b	MetricX-23-c	MetricX-23	XCOMET-Ensemble	XCOMET-XL	XCOMET-XXL	XLsim	eBLEU	embed_llama	mt0regressor	partokengram_F	tokengram_F	avg
Ambiguity	146	84	71	90	84	86	88	**97**	88	42	77	60	42	77	42	39	75	85	82	46	56	38	59	58	87	86	**95**	14	84	90	92	50	88	**95**	88	82	88	62	**93**	70	62	81	62	88	73
Coordination & ellipsis	836	69	61	**80**	78	69	61	74	63	73	74	75	74	60	74	71	78	79	80	48	77	38	75	74	76	**83**	80	58	75	62	63	33	76	**81**	72	79	**81**	64	75	61	56	72	49	62	69
False friends	225	65	63	69	**74**	71	71	67	65	71	61	75	71	65	71	73	64	80	66	46	76	**82**	64	52	66	76	63	68	74	83	**87**	30	60	82	62	65	59	53	76	66	55	56	52	68	67
Function word	200	90	78	79	90	82	74	**93**	74	92	**94**	92	90	82	90	78	86	72	85	36	**95**	62	**94**	94	94	90	67	84	86	82	76	60	86	86	84	**94**	86	64	76	70	52	76	60	76	80
MWE	829	79	72	87	**90**	86	77	85	74	74	86	79	73	66	73	71	88	89	85	45	83	37	78	78	**90**	88	89	47	89	76	78	18	86	**92**	85	90	**93**	76	81	69	69	72	54	78	76
Named entity & termin.	1272	57	54	**66**	64	62	61	**67**	62	54	51	49	55	56	54	55	61	65	58	44	54	46	60	54	68	62	**80**	56	62	60	63	30	72	**74**	71	62	60	60	71	60	48	54	49	60	59
Negation	174	87	83	89	**92**	**92**	84	84	86	90	90	84	89	76	89	92	83	88	78	47	84	40	90	90	**96**	89	91	90	**91**	**91**	**91**	58	85	**91**	80	84	85	79	80	86	67	88	52	86	83
Non-verbal agreement	372	74	72	81	**88**	78	70	83	75	82	90	84	81	83	81	78	**95**	93	91	45	**95**	43	93	91	93	82	94	71	87	76	69	57	90	**95**	87	**94**	91	88	73	72	65	84	54	69	80
Punctuation	336	69	73	74	70	66	72	**77**	69	**82**	72	76	**82**	60	**82**	65	73	66	73	42	60	44	81	**82**	**81**	75	70	75	65	**74**	72	28	**74**	70	67	65	65	46	68	**74**	49	57	44	73	68
Subordination	994	78	74	81	84	78	75	**86**	74	**89**	**88**	**88**	**89**	76	**89**	80	83	85	83	42	**89**	43	83	83	86	83	84	76	83	79	78	16	83	84	83	**89**	85	79	75	75	71	75	56	76	78
Verb tense/aspect/mood	3081	68	62	**70**	**70**	**69**	69	67	64	71	72	75	75	71	75	61	82	**84**	**85**	43	83	52	62	65	75	71	77	61	70	70	74	46	77	77	78	**83**	76	71	69	72	55	71	62	69	70
Verb valency	480	73	64	**82**	79	74	70	79	69	77	81	85	77	62	77	71	86	80	86	42	87	52	79	80	77	**87**	**88**	66	77	70	70	36	**84**	81	**85**	**85**	80	75	79	66	69	65	70	70	74
macro avg.	8945	74	69	**79**	**80**	76	73	**80**	72	75	78	77	75	70	75	70	80	80	79	44	78	48	76	75	**82**	**81**	**81**	64	79	76	76	38	80	**84**	79	81	79	68	76	70	60	71	55	73	73
micro avg.	8945	70	65	**75**	**76**	72	69	74	68	73	74	75	74	68	74	67	79	**81**	80	44	78	47	71	71	78	77	**81**	62	74	71	73	36	79	**81**	78	**81**	77	70	73	69	59	69	56	70	71

Table 2: Accuracy of the metrics (%) with regard to the 12 linguistically motivated categories for English-German

		baselines								QE as a metric																	ref. based metrics												
ling. category	#	BERTscore	BLEU	BLEURT-20	COMET	YiSi-1	chrF	prismRef	spBLEU	CometKiwi-XL	CometKiwi-XXL	CometKiwi	GEMBA-MQM	KG-BERT	MS-COMET-QE	MetricX-23-QE-b	MetricX-23-QE-c	MetricX-23-QE	Random-sysname	XCOMET-QE-Ensemble	XLsimQE	cometoid22-wmt21	cometoid22-wmt22	cometoid22-wmt23	prismSrc2	MaTESe	MetricX-23-b	MetricX-23-c	MetricX-23	XCOMET-Ensemble	XCOMET-XL	XCOMET-XXL	XLsim	eBLEU	embed_llama	partokengram_F	tokengram_F	avg
Ambiguity	66	80	67	**100**	**97**	88	67	83	73	**100**	94	**100**	**100**	**100**	95	**100**	**100**	**100**	39	**100**	30	**100**	**100**	**100**	56	**98**	**100**	**100**	**100**	**100**	**100**	**100**	88	77	44	59	70	86
Coordination & ellipsis	203	**84**	73	68	73	76	67	77	69	73	64	68	38	68	**76**	62	**79**	64	45	64	41	70	67	69	55	21	61	69	56	67	64	50	74	**80**	60	57	67	64
False friends	6	**100**	17	**100**	**100**	**100**	17	83	50	0	0	0	**100**	0	0	0	**100**	17	17	0	17	33	33	**100**	0	17	**100**	**100**	**100**	33	33	**100**	83	**100**	50	83	33	50
Function word	116	59	**69**	**69**	59	53	38	48	**72**	53	55	50	51	50	36	66	55	48	44	**72**	39	**73**	64	54	28	9	**63**	57	**60**	**60**	**60**	22	49	**61**	52	53	38	53
MWE	122	70	64	**79**	**80**	74	64	**79**	67	72	70	77	57	77	77	75	**88**	75	40	67	35	75	70	77	23	57	87	**91**	85	70	73	69	75	66	59	59	64	69
Named entity & termin.	243	79	69	**95**	91	79	76	85	76	84	62	59	54	59	69	63	73	73	49	64	53	85	86	**95**	31	51	**89**	86	86	**91**	**91**	71	84	75	65	58	75	73
Negation	34	74	74	79	68	74	**85**	68	74	68	85	85	91	85	41	71	**100**	9	59	91	9	62	29	59	91	0	**94**	71	**94**	79	71	74	65	76	76	85	85	69
Non-verbal agreement	61	69	56	**82**	77	67	67	61	57	84	**100**	54	72	54	74	72	**98**	66	41	77	26	95	**100**	85	62	43	**92**	**92**	**92**	85	79	31	64	61	66	69	69	70
Punctuation	121	42	46	53	68	**86**	41	73	55	66	29	65	27	65	**87**	55	66	43	41	61	44	66	67	36	56	1	60	71	52	66	57	12	63	66	**84**	45	45	54
Subordination	499	62	62	61	**70**	**67**	66	63	63	72	66	66	41	66	**75**	65	66	61	45	**77**	32	68	61	65	37	18	60	**84**	58	80	70	52	63	54	57	65	65	61
Verb tense/aspect/mood	135	79	62	76	**85**	76	**84**	81	74	79	79	73	61	73	64	67	65	66	39	74	60	70	59	**85**	38	50	90	**94**	90	85	**92**	74	79	70	73	82	82	73
Verb valency	121	70	67	70	**75**	**76**	67	56	68	71	**88**	79	66	79	71	76	85	76	36	76	26	63	61	65	25	45	79	81	78	72	70	65	**84**	76	64	69	69	68
macro avg.	1727	72	60	**78**	**79**	**76**	61	71	66	69	66	65	63	65	64	64	**81**	58	41	69	34	72	66	74	42	34	**81**	**83**	79	74	72	60	73	72	62	65	64	66
micro avg.	1727	69	64	72	**76**	73	65	70	67	**74**	67	68	51	68	**71**	67	**74**	64	44	**72**	39	**73**	68	**71**	40	32	73	**82**	71	77	74	55	71	67	62	63	66	66

Table 3: Accuracy of the metrics (%) with regard to the linguistically motivated categories for English-Russian

			baselines								QE as a metric														ref. based metrics																
ling. category	ling. phenomenon	#	BERTscore	BLEU	BLEURT-20	COMET	YiSi-1	chrF	prismRef	spBLEU	CometKiwi-XL	CometKiwi-XXL	CometKiwi	GEMBA-MQM	MS-COMET-QE	MetricX-23-QE-b	MetricX-23-QE-c	MetricX-23-QE	Random-sysname	XCOMET-QE-Ensemble	XLsimQE	cometoid22-wmt21	cometoid22-wmt22	prismSrc2	KG-BERT	MEE4	MEE4_stsb_xlm	MaTESe	MetricX-23-b	MetricX-23-c	MetricX-23	XCOMET-Ensemble	XCOMET-XL	XCOMET-XXL	XLsim	cometoid22-wmt23	eBLEU	embed_llama	partokengram_F	tokengram_F	avg
Ambiguity	Lexical ambiguity	129	91	75	95	**97**	89	88	95	83	81	**97**	88	79	82	81	91	84	50	91	51	84	87	63	88	84	**97**	83	91	95	95	95	93	96	94	**98**	78	73	49	86	85
	Structural ambiguity	169	85	69	84	84	**90**	75	88	76	75	84	82	75	82	76	82	77	43	84	68	**87**	84	69	82	82	80	67	83	86	85	83	86	**89**	76	86	71	44	47	75	77
Composition	Compound	129	91	64	**93**	90	**92**	71	90	71	80	85	77	79	70	81	**91**	80	43	84	29	74	74	64	77	91	90	70	87	**95**	87	88	86	89	80	86	71	61	57	72	78
	Phrasal verb	123	**89**	67	82	85	85	77	81	72	87	**89**	84	82	80	**89**	**89**	83	46	87	50	87	87	76	84	78	82	67	82	85	83	86	81	76	77	**89**	77	69	66	79	79
Coordination & ellipsis	Gapping	51	86	69	88	88	84	75	**90**	67	**96**	**96**	92	75	78	88	94	90	49	94	51	82	86	80	92	84	78	67	**96**	90	94	94	94	86	73	94	80	65	57	73	82
	Right node raising	67	81	69	72	69	**84**	73	**84**	69	75	57	**84**	76	61	75	58	69	52	66	42	55	52	76	**84**	79	69	45	81	78	73	73	72	64	70	75	72	70	55	73	69
	Sluicing	128	80	81	80	79	81	**84**	81	80	82	74	**85**	69	68	77	77	78	45	79	73	79	77	78	**85**	82	77	49	79	82	80	80	80	69	84	77	73	69	59	**84**	76
	Stripping	70	80	71	**84**	76	**84**	74	77	73	**84**	79	74	59	64	73	81	80	37	80	57	81	77	83	74	77	81	40	74	81	79	79	74	64	73	**84**	71	66	59	76	73
False friends	False friends	90	90	64	**91**	**92**	88	76	84	69	82	82	82	77	72	**89**	**90**	88	47	80	51	87	83	86	82	87	80	73	**96**	**96**	**96**	84	84	74	74	93	72	68	51	72	80
Function word	Focus particle	64	**88**	78	84	80	**88**	75	83	78	77	81	81	84	70	86	78	83	47	80	55	78	81	**88**	81	88	80	48	86	84	**89**	81	72	70	81	81	80	73	64	78	78
	Modal particle	166	84	77	84	89	89	77	**90**	81	88	92	72	72	69	84	**95**	86	43	**94**	55	92	93	69	72	86	85	51	89	89	88	91	89	78	84	**94**	81	70	54	76	80
	Question tag	356	80	70	**82**	**81**	78	71	81	69	83	**86**	76	78	67	77	83	77	49	83	63	82	**85**	85	77	74	71	61	86	81	84	86	82	70	78	**92**	78	63	54	74	76
LDD & interrogatives	Extended adjective construction	320	85	78	85	**90**	86	80	88	79	76	80	**88**	70	**88**	85	83	82	49	78	62	81	83	74	88	84	78	59	82	85	82	85	81	78	81	**92**	81	67	53	82	79
	Extraposition	92	77	75	**79**	**80**	77	**79**	78	74	68	68	70	54	66	62	67	66	48	**74**	50	**74**	**74**	49	70	**84**	78	37	76	75	80	73	65	42	75	76	82	74	51	77	69
	Multiple connectors	87	79	79	72	**84**	**84**	77	76	78	**80**	75	76	61	78	71	75	74	48	72	66	72	56	69	76	**78**	71	24	70	70	62	74	77	46	68	**79**	71	67	52	77	70
	Pied-piping	162	**90**	73	**89**	**89**	**88**	73	**89**	71	**87**	**86**	85	76	73	84	85	85	38	85	71	85	81	**87**	85	83	78	56	**89**	86	**88**	86	85	75	78	87	77	67	60	73	80
	Polar question	51	**84**	63	76	76	76	53	76	61	84	82	84	75	86	**88**	78	**88**	43	78	57	84	84	73	84	71	65	47	78	80	78	80	80	59	67	**90**	63	67	53	57	73
	Scrambling	144	81	72	87	87	85	78	**90**	72	81	89	85	78	83	**91**	**91**	86	41	87	58	81	81	67	85	75	77	69	83	**90**	88	88	86	84	79	85	72	61	56	78	79
	Topicalization	61	90	70	**97**	90	89	72	89	70	90	82	**97**	90	89	**97**	92	93	41	85	72	90	92	74	**97**	87	90	59	92	92	90	90	93	82	77	93	75	80	43	75	83
	Wh-movement	97	78	68	85	**87**	79	74	85	69	82	84	75	82	71	**86**	**86**	**85**	49	76	49	78	78	69	75	78	65	52	76	78	78	85	**87**	76	71	**87**	67	76	54	74	75
MWE	Collocation	190	80	71	**86**	**85**	**86**	76	81	74	77	85	76	84	72	87	**92**	82	45	84	55	75	74	67	76	79	79	70	84	**91**	86	86	86	74	80	83	72	62	51	76	77
	Idiom	133	84	72	83	89	88	78	**95**	72	62	77	59	64	32	73	**80**	71	41	78	43	58	59	41	59	80	93	69	80	92	86	80	84	88	75	**96**	81	75	51	77	73
	Prepositional MWE	146	82	77	86	82	85	81	**88**	77	78	85	**87**	71	67	80	84	84	52	**86**	50	79	79	72	87	86	78	66	88	84	**92**	87	79	81	74	81	78	67	57	81	78
	Verbal MWE	141	**82**	67	80	80	81	70	**82**	70	78	83	84	72	61	79	**92**	81	42	83	62	77	79	78	82	74	74	62	**86**	74	**86**	82	84	80	70	82	71	63	57	70	75
Named entity & termin.	Date	203	67	54	67	70	67	61	**77**	61	78	**80**	70	68	64	68	76	70	43	75	67	67	66	75	70	66	66	62	73	**76**	**76**	74	**76**	70	70	72	57	54	50	60	68
	Domainspecific term	214	**73**	60	70	69	**73**	70	64	66	58	63	64	51	63	66	65	62	43	61	62	66	**69**	49	65	70	72	49	72	**76**	71	69	**76**	66	71	69	60	51	50	71	64
	Location	181	72	69	70	**81**	**82**	69	72	77	54	55	60	55	61	58	66	65	48	56	42	62	62	**74**	60	**74**	69	57	65	70	68	63	72	57	70	70	58	60	52	72	64
	Measuring unit	203	79	69	65	72	80	73	76	**84**	66	65	63	62	55	64	71	63	42	70	58	62	66	**77**	64	75	69	61	72	69	68	74	74	66	63	70	**79**	61	54	76	68
	Proper name	60	77	73	83	83	**88**	77	87	75	78	82	83	82	80	82	75	85	47	83	78	85	**87**	83	83	83	83	72	85	85	85	**88**	85	83	82	83	73	65	48	80	79
Negation	Negation	76	79	78	91	**93**	84	82	84	76	91	91	93	75	84	91	**95**	86	50	93	55	91	89	92	93	86	84	51	91	**95**	92	**95**	91	79	76	91	75	59	45	82	82
Non-verbal agreement	Coreference	251	76	71	**89**	87	77	75	77	72	86	**89**	81	72	83	86	**88**	85	43	87	67	84	85	74	81	74	73	62	87	**89**	**88**	87	81	80	64	86	72	57	69	74	78
	External possessor	104	82	88	75	78	81	84	80	**90**	62	65	65	59	**68**	62	**68**	**67**	38	66	59	58	62	**67**	65	83	79	48	67	79	65	69	66	58	72	70	**88**	69	83	83	70
	Internal possessor	64	81	**88**	84	83	80	78	**88**	**88**	73	72	78	58	75	72	73	77	47	75	64	83	80	**89**	77	83	78	66	78	75	78	78	77	66	**88**	78	77	59	80	80	76
Punctuation	Comma	46	80	76	61	67	**83**	74	78	**83**	48	52	61	46	63	46	54	54	35	59	65	70	67	**85**	61	**78**	74	48	52	43	50	57	63	37	61	63	70	65	74	74	63
	Quotation marks	247	74	74	70	69	68	63	**81**	74	64	64	64	62	58	61	62	61	40	66	68	65	66	**86**	64	**76**	66	48	64	57	64	67	66	45	70	67	64	66	66	66	65
Subordination	Adverbial clause	87	62	56	**75**	**75**	69	56	**75**	54	**82**	71	78	54	63	77	69	77	44	69	51	74	74	68	**78**	72	68	49	68	74	64	74	67	49	68	**77**	62	66	56	56	66
	Cleft sentence	109	75	**83**	74	74	69	78	64	80	73	72	72	49	59	68	71	**76**	50	74	69	65	67	55	72	71	68	45	69	69	67	73	72	53	66	71	**79**	72	**78**	**78**	69
	Free relative clause	70	77	64	77	**84**	70	67	79	66	71	**90**	**90**	63	69	71	79	71	46	80	60	76	73	59	**90**	74	79	59	71	74	73	81	74	69	77	80	70	59	67	67	72
	Indirect speech	119	75	68	**79**	**79**	73	73	**80**	66	84	86	74	71	68	82	79	80	43	**89**	43	80	81	79	74	81	80	56	84	82	79	**86**	83	76	69	82	72	61	74	74	75
	Infinitive clause	64	80	80	84	86	84	81	**88**	77	78	78	78	61	62	80	73	**81**	34	73	38	73	72	64	78	**88**	78	56	80	81	80	80	75	78	67	80	77	55	80	80	74
	Object clause	54	**76**	61	56	74	74	59	74	67	69	69	80	76	72	74	**87**	78	46	69	56	76	74	72	**80**	74	72	31	70	74	74	70	57	37	72	78	72	59	63	63	68
	Pseudo-cleft sentence	25	80	60	72	60	72	64	**92**	60	76	72	80	76	20	**92**	84	84	36	84	12	72	68	60	80	80	72	52	**88**	**88**	84	84	**88**	68	64	84	80	44	64	64	70
	Relative clause	71	76	68	**82**	76	80	70	80	72	82	75	79	62	62	79	77	77	35	**83**	48	68	76	63	79	76	72	46	86	83	**92**	82	82	66	73	80	73	63	72	72	73

Accuracy of the metrics(%) with regards to the linguistically-motivated phenomena for German-English - Continued on next page

			baselines								QE as a metric														ref. based metrics																
ling. category	ling. phenomenon	#	BERTscore	BLEU	BLEURT-20	COMET	YiSi-1	chrF	prismRef	spBLEU	CometKiwi-XL	CometKiwi-XXL	CometKiwi	GEMBA-MQM	MS-COMET-QE	MetricX-23-QE-b	MetricX-23-QE-c	MetricX-23-QE	Random-sysname	XCOMET-QE-Ensemble	XLsimQE	cometoid22-wmt21	cometoid22-wmt22	prismSrc2	KG-BERT	MEE4	MEE4_stsb_xlm	MaTESe	MetricX-23-b	MetricX-23-c	MetricX-23	XCOMET-Ensemble	XCOMET-XL	XCOMET-XXL	XLsim	cometoid22-wmt23	eBLEU	embed_llama	partokengram_F	tokengram_F	avg
	Subject clause	80	82	78	80	**86**	84	74	**86**	76	**88**	**86**	84	76	80	85	85	84	48	84	71	**86**	**88**	82	84	82	81	49	85	81	**86**	84	80	74	75	**88**	78	59	74	74	79
Verb tense/aspect/mood	Conditional	50	90	68	**92**	**92**	88	66	84	64	90	**96**	**96**	92	86	92	88	88	34	88	64	80	80	82	**96**	74	70	52	88	94	86	92	92	86	78	90	66	40	62	62	80
	Ditransitive - future I	121	89	69	88	87	**91**	74	**90**	71	95	**97**	94	**96**	93	93	95	90	53	92	56	80	81	91	**94**	79	74	63	91	**95**	89	93	92	89	81	93	74	57	74	74	84
	Ditransitive - future I subj. II	84	**96**	62	86	90	94	75	89	68	**100**	94	**100**	**100**	95	**100**	98	95	35	95	52	75	80	94	**100**	79	82	57	80	92	86	96	88	88	63	95	65	57	75	75	83
	Ditransitive - future II	97	**98**	65	84	90	96	72	84	66	93	94	97	94	96	88	**99**	79	41	85	66	77	77	86	97	72	73	68	84	91	87	92	90	82	68	**99**	67	53	71	71	81
	Ditransitive - future II subj. II	88	90	69	88	**94**	**94**	77	91	73	98	98	98	95	92	97	**100**	89	39	95	67	91	90	92	**98**	78	78	81	93	**97**	93	93	92	91	76	93	80	55	76	76	86
	Ditransitive - perfect	72	89	64	78	88	**90**	72	79	62	88	83	83	75	79	**89**	76	83	50	75	58	72	75	72	83	68	75	42	75	81	69	75	71	60	74	**89**	76	69	72	72	75
	Ditransitive - pluperfect	86	**91**	69	72	**90**	**90**	86	78	69	55	66	64	64	72	52	**74**	43	51	55	52	49	52	**76**	64	78	76	43	73	72	71	69	66	45	71	**97**	74	67	87	87	69
	Ditransitive - pluperfect subj. II	107	82	68	**91**	87	83	80	88	70	89	89	84	93	79	93	**95**	82	40	84	55	69	64	93	84	82	79	64	**93**	**93**	**93**	92	90	91	78	92	74	64	79	79	81
	Ditransitive - present	90	87	58	87	86	**89**	77	83	64	88	90	89	**96**	82	92	**97**	92	53	82	51	69	70	69	89	80	81	69	93	88	90	**97**	93	90	79	89	70	69	79	79	81
	Ditransitive - preterite	117	88	62	**93**	**92**	89	82	**93**	66	93	88	92	90	84	**95**	**95**	**94**	48	89	59	83	84	81	92	89	85	81	96	97	96	95	**98**	88	76	**97**	73	68	85	85	85
	Ditransitive - preterite subj. II	110	83	76	93	92	89	83	**95**	76	79	88	84	86	76	90	85	**92**	45	85	65	78	79	72	84	92	89	75	**98**	88	96	92	95	85	77	95	77	70	85	85	84
	Imperative	98	84	59	**94**	93	91	74	**95**	63	**94**	90	**95**	**94**	**94**	**94**	**94**	93	44	**94**	57	**94**	93	84	**95**	90	84	73	**95**	92	90	**95**	91	92	79	**96**	76	59	77	77	85
	Intransitive - future I	32	91	66	94	91	91	75	**100**	75	97	**100**	88	97	78	**100**	97	97	47	97	66	69	81	81	88	94	91	62	**100**	94	**100**	97	88	94	75	94	75	72	78	78	86
	Intransitive - future I subj. II	56	84	61	93	88	91	73	**96**	70	82	82	77	80	62	77	**86**	80	36	**86**	32	66	68	59	77	89	**95**	59	89	**95**	91	84	80	80	77	**95**	70	55	73	73	77
	Intransitive - future II	62	92	69	89	95	**97**	89	**97**	76	92	76	92	69	82	90	**97**	87	44	85	73	71	77	68	92	95	**98**	55	85	94	82	85	82	77	89	**98**	73	52	87	87	83
	Intransitive - future II subj. II	94	88	65	**90**	82	**90**	72	84	63	82	82	83	**88**	71	85	86	82	44	83	54	80	79	73	83	80	83	62	83	**88**	84	83	83	80	79	84	65	54	71	71	77
	Intransitive - perfect	61	85	70	82	84	**89**	82	**89**	69	70	69	77	85	64	**93**	85	82	39	67	61	74	75	61	77	82	**92**	34	77	80	74	69	62	64	80	**92**	87	72	82	82	75
	Intransitive - pluperfect	85	**98**	86	**96**	95	95	88	**98**	82	**92**	86	**93**	**92**	73	88	91	88	44	87	62	79	82	61	93	96	**99**	74	92	95	94	88	89	91	94	**98**	85	62	91	91	87
	Intransitive - pluperfect subj. II	79	91	75	92	96	**99**	78	**97**	72	92	87	94	**97**	86	**97**	95	**96**	43	87	48	77	78	84	94	95	94	59	90	**96**	94	**97**	95	94	78	**97**	78	57	81	81	85
	Intransitive - present	54	72	50	85	83	74	61	**87**	57	81	74	**91**	85	85	87	78	74	46	74	76	83	89	74	91	76	74	46	87	91	83	81	80	85	70	**93**	67	63	63	63	76
	Intransitive - preterite	46	78	43	83	**85**	80	52	80	63	83	89	**93**	80	85	72	83	78	50	89	52	91	89	87	**93**	87	83	67	91	91	91	83	72	87	70	85	74	61	57	57	77
	Intransitive - preterite subj. II	100	89	69	88	92	88	77	**95**	76	85	84	**89**	81	87	85	83	79	44	84	52	68	72	72	89	**93**	89	64	85	89	86	88	79	80	81	**94**	88	69	81	81	81
	Modal - future I	42	**98**	**98**	74	93	**98**	95	95	**98**	**93**	83	79	57	69	74	76	67	50	76	62	33	38	57	79	90	98	29	67	74	74	86	81	55	98	98	**100**	83	95	95	78
	Modal - future I subj. II	86	86	77	81	78	**88**	76	**90**	79	85	79	78	65	71	86	86	81	50	85	56	44	42	**94**	78	83	**90**	31	65	79	67	84	74	72	80	**91**	79	57	77	77	75
	Modal - perfect	149	91	83	77	87	90	84	**93**	83	**85**	81	83	56	52	83	82	76	37	79	61	44	46	73	83	81	89	21	52	68	56	77	66	62	86	**98**	83	66	85	85	73
	Modal - pluperfect	75	88	80	81	92	92	84	**95**	77	79	76	80	75	81	76	**85**	67	37	73	77	83	80	84	80	85	92	47	75	85	80	77	79	67	89	**95**	81	76	83	83	79
	Modal - pluperfect subj. II	61	85	57	89	80	85	74	**95**	66	80	85	87	84	89	72	84	74	46	87	59	69	74	**92**	87	85	89	46	75	82	74	**90**	84	82	72	80	84	70	74	74	78
	Modal - present	30	93	67	90	90	93	87	**100**	73	90	83	90	77	90	**97**	**97**	**97**	43	90	40	73	80	**97**	90	93	93	53	**100**	97	**100**	97	93	97	90	93	90	63	87	87	86
	Modal - preterite	72	85	78	82	88	88	76	**94**	78	**94**	83	85	93	89	88	78	86	46	90	57	72	69	83	86	88	86	44	75	78	78	83	82	68	79	**89**	86	72	76	76	80
	Modal - preterite subj. II	30	**100**	**100**	80	93	**100**	93	93	97	87	87	**90**	87	**90**	87	87	**90**	30	80	73	63	63	73	90	93	**100**	40	70	83	80	77	67	47	87	**100**	97	77	93	93	83
	Modal negated - future I	43	95	91	88	95	**98**	93	**98**	91	91	95	**98**	88	93	93	93	95	47	91	77	56	58	77	**98**	88	95	42	79	91	81	91	84	65	95	95	91	77	93	93	86
	Modal negated - future I subj. II	73	78	60	77	74	78	73	**82**	59	85	67	84	68	88	**89**	75	75	48	71	63	59	56	78	84	79	78	33	59	71	64	75	70	64	84	**92**	63	52	75	75	71
	Modal negated - perfect	126	81	67	71	82	80	77	**87**	71	79	71	75	64	80	79	75	77	46	67	**83**	54	54	56	75	84	84	33	61	66	65	67	58	52	80	**94**	75	63	78	78	71
	Modal negated - pluperfect	126	86	77	83	85	86	87	**90**	79	72	68	75	71	**79**	75	72	70	39	63	76	72	71	67	75	84	**89**	38	74	84	81	64	67	50	**89**	**89**	79	65	87	87	75
	Modal negated - pluperfect subj. II	81	78	70	80	83	80	75	**85**	57	84	77	75	**91**	81	84	77	80	42	78	69	80	81	**90**	75	**84**	80	38	77	80	79	79	72	65	75	81	68	54	70	70	75
	Modal negated - present	33	91	64	**97**	91	91	91	94	55	94	91	97	88	97	97	**100**	94	52	94	48	82	88	**100**	97	94	91	61	91	94	91	97	**100**	91	79	**100**	67	67	91	91	87
	Modal negated - preterite	61	84	72	90	89	**92**	87	90	69	89	74	75	67	89	87	84	72	38	77	38	75	75	**90**	75	89	85	49	77	85	87	84	80	70	84	**90**	75	62	87	87	78
	Modal negated - preterite subj. II	77	77	66	82	81	78	70	**88**	68	**91**	79	86	61	79	**90**	75	77	34	84	56	70	74	82	86	84	78	57	83	**91**	84	88	81	74	69	87	74	73	69	69	76
	Progressive	76	82	66	79	**86**	79	71	82	63	76	**83**	79	78	67	75	76	74	45	82	59	72	70	68	79	68	75	57	80	**83**	78	**83**	79	74	75	82	70	57	68	68	73
	Reflexive - future I	85	**86**	76	**86**	**87**	85	79	**87**	74	72	**89**	78	85	71	87	85	**89**	36	87	54	68	69	69	78	84	86	62	87	91	87	88	88	84	78	**94**	80	45	79	79	79
	Reflexive - future I subj. II	96	85	77	69	77	86	73	**90**	72	69	79	**83**	**82**	75	72	**82**	66	42	71	66	78	73	72	83	76	80	38	76	78	75	79	79	69	74	**89**	70	62	73	73	74
	Reflexive - future II	116	94	77	83	92	**96**	79	91	75	72	80	84	**88**	78	76	85	69	49	79	53	72	76	81	84	82	86	54	77	81	76	85	84	72	77	**97**	71	53	78	78	78

Accuracy of the metrics(%) with regards to the linguistically-motivated phenomena for German-English - Continued on next page

			baselines								QE as a metric															ref. based metrics																
ling. category	ling. phenomenon	#	BERTscore	BLEU	BLEURT-20	COMET	YiSi-1	chrF	prismRef	spBLEU	CometKiwi-XL	CometKiwi-XXL	CometKiwi	GEMBA-MQM	MS-COMET-QE	MetricX-23-QE-b	MetricX-23-QE-c	MetricX-23-QE	Random-sysname	XCOMET-QE-Ensemble	XLsimQE	cometoid22-wmt21	cometoid22-wmt22	prismSrc2	KG-BERT	MEE4	MEE4_stsb_xlm	MaTESe	MetricX-23-b	MetricX-23-c	MetricX-23	XCOMET-Ensemble	XCOMET-XL	XCOMET-XXL	XLsim	cometoid22-wmt23	eBLEU	embed_llama	partokengram_F	tokengram_F	avg
	Reflexive - future II subj. II	107	90	67	84	**93**	92	71	**93**	70	89	**93**	91	**93**	76	84	88	79	41	92	63	84	92	83	91	75	82	55	86	84	88	**94**	**93**	82	69	**93**	69	57	73	73	81
	Reflexive - perfect	188	**85**	61	83	**85**	80	60	**85**	64	82	88	**90**	80	78	83	**90**	80	46	88	58	79	84	84	90	76	74	63	84	**91**	88	86	83	82	69	**91**	70	57	60	60	77
	Reflexive - pluperfect	109	**96**	74	76	91	94	67	86	71	70	77	80	**84**	75	66	**84**	62	43	73	55	72	71	81	80	80	81	55	69	83	67	79	74	70	77	**96**	72	75	72	72	75
	Reflexive - pluperfect subj. II	90	90	68	87	90	87	74	**92**	74	78	87	87	**92**	88	89	83	78	38	86	64	89	84	88	87	84	86	58	79	82	79	84	81	84	76	**92**	83	69	76	76	81
	Reflexive - present	125	82	55	**90**	86	82	72	88	66	79	84	88	81	**91**	89	**92**	87	46	85	58	89	86	**91**	88	81	78	62	87	90	84	**92**	87	86	69	88	70	58	70	70	80
	Reflexive - preterite	117	**92**	75	86	85	90	77	**92**	77	71	68	74	**89**	83	80	80	73	39	71	61	81	80	78	74	**92**	82	64	90	**93**	86	77	82	79	71	85	85	79	80	80	79
	Reflexive - preterite subj. II	124	94	89	90	93	**97**	81	**96**	85	86	81	90	**96**	88	93	86	93	44	85	55	82	82	88	90	95	90	77	**98**	**98**	**98**	88	96	92	77	97	85	73	83	83	87
	Transitive - future I	43	98	84	95	95	**100**	86	98	86	84	**93**	77	88	91	84	81	79	44	**93**	51	84	84	**93**	77	95	91	60	93	**98**	93	93	86	86	86	95	65	56	86	86	85
	Transitive - future I subj. II	37	97	81	86	95	**100**	89	95	81	95	95	84	81	**100**	95	97	89	43	92	68	70	95	**100**	84	89	89	62	86	89	84	89	89	73	84	**97**	81	54	84	84	85
	Transitive - future II	33	97	82	94	94	**100**	91	94	79	91	94	97	91	85	94	97	82	45	94	67	76	85	**100**	**97**	85	**97**	61	**97**	**97**	**97**	94	88	79	94	**97**	73	36	94	94	87
	Transitive - future II subj. II	50	**96**	62	78	80	86	82	68	74	92	**98**	70	92	66	96	**98**	**98**	56	96	54	90	88	86	70	72	70	50	82	76	82	80	80	80	82	**96**	68	56	82	82	79
	Transitive - perfect	99	88	73	76	91	**95**	85	**96**	77	70	77	72	84	82	82	83	72	43	73	60	62	69	**86**	72	88	94	58	83	84	84	81	78	73	85	**96**	80	72	86	86	79
	Transitive - pluperfect	22	82	41	86	86	86	59	**95**	50	77	91	82	**95**	**95**	91	91	77	27	91	45	68	68	91	86	82	64	77	82	82	86	**91**	**91**	77	68	**91**	64	32	55	55	75
	Transitive - pluperfect subj. II	39	90	64	**92**	90	90	72	**92**	69	79	87	92	**95**	**95**	92	90	92	49	82	51	82	85	85	92	97	87	87	97	**100**	97	87	90	90	85	95	85	59	72	72	84
	Transitive - present	33	73	52	85	**91**	82	67	**91**	70	82	88	91	**100**	91	85	97	82	48	85	70	88	91	70	91	94	88	88	**97**	94	**97**	88	85	94	79	**97**	85	64	76	76	83
	Transitive - preterite	57	82	39	**84**	79	77	56	77	54	**84**	81	81	63	72	77	77	75	40	79	58	81	81	74	81	77	70	58	77	79	75	82	75	**84**	68	**84**	74	53	58	58	72
	Transitive - preterite subj. II	97	81	48	**88**	**88**	**89**	59	**88**	54	82	87	92	**94**	84	85	**94**	77	41	88	65	82	85	79	92	78	78	77	92	91	**95**	93	**95**	92	74	88	64	62	62	62	80
Verb valency	Case government	80	89	74	86	**91**	85	72	**90**	70	**84**	74	**84**	60	70	**82**	78	78	57	76	50	79	80	71	84	84	70	57	86	86	86	86	89	74	80	**92**	70	51	75	75	77
	Mediopassive voice	50	94	72	92	90	**96**	78	94	72	90	84	88	84	86	90	**96**	88	54	94	40	88	94	72	88	90	88	68	90	90	**94**	90	90	84	92	**94**	86	58	80	80	84
	Passive voice	33	**55**	**55**	**55**	**55**	**55**	**55**	**55**	**55**	39	39	42	18	27	**45**	39	30	27	39	27	42	39	27	42	**55**	**55**	36	52	**55**	39	48	36	33	**55**	**55**	**55**	52	**55**	**55**	45
macro avg.		10402	84	69	82	84	**85**	75	**86**	71	80	81	81	76	75	81	**83**	79	43	80	57	74	75	76	81	81	80	56	81	84	81	83	80	73	76	**87**	74	62	69	75	76
micro avg.		10402	83	70	82	**84**	**84**	75	**85**	72	79	80	80	75	74	80	**83**	78	44	80	58	75	76	75	80	81	79	57	81	83	81	82	81	74	76	**87**	74	63	67	75	76

Table 4: Accuracy of the metrics(%) with regard to the linguistically-motivated phenomena for German-English

ling. phenomenon	#	baselines								QE as a metric																				ref. based metrics																	
		BERTscore	BLEU	BLEURT-20	COMET	YiSi-1	chrF	prismRef	spBLEU	Calibri-COMET22-QE	CometKiwi-XL	CometKiwi-XXL	CometKiwi	GEMBA-MQM	KG-BERT	MS-COMET-QE	MetricX-23-QE-b	MetricX-23-QE-c	MetricX-23-QE	Random-sysname	XCOMET-QE-Ensemble	XLsimQE	cometoid22-wmt21	cometoid22-wmt22	cometoid22-wmt23	mltsamp-bleurt0p2p1-qe	mltsamp-bleurtxv1p-qe	prismSrc2	Calibri-COMET22	MEE4	MEE4_stsb_xlm	MaTESe	MetricX-23-b	MetricX-23-c	MetricX-23	XCOMET-Ensemble	XCOMET-XL	XCOMET-XXL	XLsim	eBLEU	embed_llama	mt0regressor	partokengram_F	tokengram_F	avg
Ambiguity																																														
Lexical ambiguity	146	84	71	90	84	86	88	**97**	88	42	77	60	42	77	42	39	75	85	82	46	56	38	59	58	87	86	**95**	14	84	90	92	50	88	**95**	88	82	88	62	**93**	70	62	81	62	88	73
Coordination & ellipsis																																														
Gapping	163	72	59	86	84	71	67	**88**	66	75	64	79	76	84	76	78	91	94	93	42	84	56	78	75	75	**98**	91	66	83	63	67	65	76	**89**	71	83	**88**	71	74	58	55	61	51	67	74
Pseudogapping	201	81	76	**98**	92	79	67	81	73	94	97	**98**	94	83	94	79	93	98	93	45	**98**	23	96	95	**99**	93	96	63	88	74	73	42	90	96	84	**98**	**98**	84	87	75	62	76	48	68	82
Right node raising	47	83	64	87	**98**	83	72	89	70	91	87	89	91	81	91	85	91	91	85	53	89	9	91	89	**94**	89	85	87	89	68	70	51	89	**98**	87	89	91	83	81	55	64	91	55	72	80
Sluicing	169	55	**58**	**57**	50	54	54	45	**58**	37	47	53	39	19	39	**69**	49	56	57	51	57	49	39	41	44	63	54	43	44	56	59	11	57	66	53	64	67	44	57	56	57	**69**	44	55	51
Stripping	139	72	58	68	**76**	68	58	**76**	61	64	73	50	64	45	64	36	65	56	69	50	55	46	65	62	68	74	**77**	45	73	63	60	29	68	66	75	56	60	40	**78**	60	58	71	50	58	61
VP-ellipsis	117	59	47	**85**	84	68	53	74	45	88	85	84	**89**	59	**89**	84	81	82	82	52	82	25	**90**	**90**	85	85	79	67	81	49	50	6	82	78	71	**85**	81	67	74	49	44	79	51	56	70
False friends																																														
False friends	225	65	63	69	**74**	71	71	67	65	71	61	75	71	65	71	73	64	80	66	46	76	**82**	64	52	66	76	63	68	74	83	**87**	30	60	82	62	65	59	53	76	66	55	56	52	68	67
Function word																																														
Focus particle	20	45	30	**70**	60	50	35	**70**	30	85	90	90	85	50	85	50	90	30	40	40	**100**	95	50	50	50	70	65	40	60	35	35	25	80	**100**	70	95	60	45	45	35	40	80	40	35	59
Question tag	180	**96**	83	80	93	86	78	**96**	79	92	95	92	91	86	91	81	86	77	90	36	94	58	**99**	**98**	**98**	93	67	89	89	88	81	64	87	84	86	**94**	89	66	79	74	53	75	63	80	83
MWE																																														
Collocation	112	73	62	**93**	90	89	77	88	60	84	96	92	80	81	80	74	**97**	93	**97**	45	**98**	15	81	73	83	**97**	**97**	57	88	82	81	30	95	96	**99**	**99**	**99**	79	80	61	68	71	51	78	79
Compound	63	73	51	56	89	95	84	**98**	68	95	97	**100**	**100**	79	**100**	56	**100**	98	**100**	48	**100**	22	86	94	81	90	98	83	81	62	65	2	**100**	**100**	92	**100**	83	79	78	51	65	40	51	84	78
Idiom	266	86	82	**97**	**98**	93	75	93	83	89	93	97	90	**99**	90	75	98	**100**	97	44	96	23	82	84	**98**	95	98	36	**98**	82	89	15	95	97	89	96	97	86	80	80	71	81	60	76	83
Nominal MWE	288	81	72	83	**86**	80	78	75	74	51	75	54	50	26	50	71	74	78	65	42	66	61	73	72	**92**	81	73	39	86	72	70	17	72	84	76	83	**93**	71	85	65	70	71	54	79	69
Prepositional MWE	35	69	**86**	71	83	83	83	83	83	83	80	80	83	71	83	**89**	80	71	83	63	83	46	86	86	**89**	71	**89**	80	83	83	83	17	71	**89**	57	83	80	69	80	77	74	83	46	83	77
Verbal MWE	65	68	72	**88**	77	66	74	78	60	69	75	71	66	66	66	55	94	86	95	43	65	37	72	66	75	86	**98**	60	77	77	74	32	**95**	88	89	69	75	57	75	74	57	66	49	71	71
Named entity & termin.																																														
Date	234	52	55	**74**	71	64	60	68	63	67	72	60	67	75	67	77	77	81	77	46	64	62	83	82	**88**	85	84	82	63	55	65	24	**85**	79	81	75	79	69	77	69	54	79	51	61	69
Domainspecific term	312	69	56	**88**	82	79	74	**88**	66	71	46	58	74	72	73	69	66	87	65	46	64	12	83	69	89	83	**98**	61	82	68	74	53	91	**98**	96	76	71	78	96	65	42	41	47	71	71
Location	12	67	83	75	67	50	58	**100**	83	67	58	67	67	83	67	92	92	75	92	58	92	67	**100**	**100**	75	75	**100**	**100**	67	92	92	67	**100**	**100**	**100**	**100**	**100**	**100**	83	83	58	**100**	58	58	81
Measuring unit	389	54	48	53	**56**	55	**56**	**57**	53	28	37	20	28	28	28	37	51	50	40	44	27	56	28	23	53	35	**77**	36	55	58	**63**	10	57	56	53	38	35	49	**62**	49	48	42	48	55	45
Proper name	325	50	57	54	50	53	54	56	**68**	57	57	**66**	58	58	58	48	56	49	58	39	**66**	54	58	56	50	57	64	55	50	58	51	35	61	**68**	59	**67**	62	49	53	61	48	60	48	56	56
Negation																																														
Negation	174	87	83	89	**92**	**92**	84	84	86	90	90	84	89	76	89	92	83	88	78	47	84	40	90	90	**96**	89	91	90	**91**	**91**	**91**	58	85	**91**	80	84	85	79	80	86	67	88	52	86	83
Non-verbal agreement																																														
Coreference	81	85	85	**95**	80	73	86	84	88	83	77	95	83	69	83	75	95	89	91	52	**98**	74	85	81	88	93	**98**	62	79	88	79	91	91	98	96	**100**	96	85	80	79	72	79	56	85	84
Genitive	206	75	73	73	**93**	82	68	80	76	77	94	78	76	83	76	90	**97**	95	90	46	96	24	**97**	**96**	**97**	75	93	69	90	72	67	53	90	**95**	81	91	87	88	76	77	65	85	52	69	79
Possession	85	62	55	**86**	**86**	74	58	**87**	60	93	93	89	93	**95**	93	53	89	92	93	39	92	58	89	87	89	91	91	86	86	72	64	34	88	**93**	**93**	**93**	**93**	91	59	54	61	86	56	54	78
Punctuation																																														
Quotation marks	336	69	73	74	70	66	72	**77**	69	**82**	72	76	**82**	60	**82**	65	73	66	73	42	60	44	81	**82**	**81**	75	70	75	65	**74**	72	28	**74**	70	67	65	65	46	68	**74**	49	57	44	73	68
Subordination																																														
Adverbial clause	193	72	80	80	82	71	73	**83**	77	91	**93**	90	91	80	91	68	83	88	87	42	88	44	79	73	76	86	83	70	80	77	75	8	75	84	74	**91**	**92**	87	67	69	67	63	51	74	76
Cleft sentence	179	67	64	61	**75**	66	59	**75**	62	76	72	74	76	58	76	75	70	72	75	41	74	45	**78**	75	**78**	64	69	63	73	68	67	14	66	65	71	**75**	69	59	63	70	66	65	54	61	66
Contact clause	150	84	75	**95**	**94**	91	75	89	75	**97**	95	**98**	**97**	93	**97**	91	87	**98**	91	47	**98**	44	94	96	**97**	91	93	91	94	79	79	47	97	97	97	**100**	**99**	89	86	79	71	91	61	77	87
Indirect speech	38	55	42	63	63	47	47	**82**	42	84	**97**	71	79	61	79	50	68	66	76	47	82	26	82	79	76	95	87	84	50	45	47	5	79	63	79	**82**	74	74	58	45	55	58	45	50	64
Infinitive clause	85	66	55	89	**92**	87	80	88	67	**98**	**99**	95	**98**	87	**98**	89	95	95	91	44	95	75	**98**	**98**	**98**	94	93	79	91	73	75	8	92	93	91	**95**	**94**	85	88	60	66	89	49	79	83
Object clause	16	56	44	**94**	69	62	62	75	50	75	**100**	88	81	81	81	88	**100**	81	**100**	31	**100**	69	81	81	81	81	**100**	81	62	50	69	38	94	81	**100**	**100**	**100**	75	50	50	50	75	56	62	75
Pseudo-cleft sentence	73	**90**	86	68	73	85	89	**92**	84	77	**82**	**82**	74	58	74	73	78	**82**	74	38	**82**	68	68	**82**	**82**	64	70	79	75	**92**	82	12	79	75	71	82	67	73	86	79	86	88	56	86	76

Accuracy of the metrics(%) with regards to the linguistically-motivated phenomena for English-German. Continued on next page

ling. phenomenon	#	baselines								QE as a metric																			ref. based metrics																					avg
		BERTscore	BLEU	BLEURT-20	COMET	YiSi-1	chrF	prismRef	spBLEU	Calibri-COMET22-QE	CometKiwi-XL	CometKiwi-XXL	CometKiwi	GEMBA-MQM	KG-BERT	MS-COMET-QE	MetricX-23-QE-b	MetricX-23-QE-c	MetricX-23-QE	Random-sysname	XCOMET-QE-Ensemble	XLsimQE	cometoid22-wmt21	cometoid22-wmt22	cometoid22-wmt23	mltsamp-bleurt0p2p1-qe	mltsamp-bleurtxv1p-qe	prismSrc2	Calibri-COMET22	MEE4	MEE4_stsb_xlm	MaTESe	MetricX-23-b	MetricX-23-c	MetricX-23	XCOMET-Ensemble	XCOMET-XL	XCOMET-XXL	XLsim	eBLEU	embed_llama	mt0regressor	partokengram_F	tokengram_F	
Relative clause	112	91	80	88	**93**	77	83	**92**	77	**100**	84	98	**100**	79	**100**	94	94	85	81	39	94	40	80	80	94	97	91	84	93	87	90	5	92	**97**	91	92	76	74	76	88	64	62	61	81	82
Subject clause	148	**92**	**93**	91	89	90	91	89	91	89	89	89	89	80	89	88	80	86	82	43	**93**	14	86	90	89	85	91	70	89	**96**	92	11	91	89	93	93	92	86	85	92	89	83	62	93	83
Verb tense/aspect/mood																																														
cond.	106	74	77	91	88	86	72	**93**	74	97	85	94	97	86	97	82	95	91	96	37	92	38	**99**	**100**	**100**	94	97	92	87	86	85	14	**97**	89	90	91	92	84	82	76	60	80	57	73	83
Ditrans. - cond. I progr.	72	65	49	93	**96**	83	61	81	57	**100**	90	**100**	**100**	**100**	**100**	92	94	**100**	94	43	**99**	65	**99**	**99**	**100**	**100**	96	94	92	61	62	46	83	**100**	85	**99**	92	74	67	57	61	82	49	Q60	82
Ditrans. - cond. I simple	34	94	74	65	88	**97**	94	91	79	88	68	97	88	76	88	91	**100**	**100**	97	56	79	79	50	71	82	65	91	47	91	**97**	**97**	74	79	85	71	**97**	79	56	94	76	85	**97**	65	94	82
Ditrans. - cond. II progr.	51	75	78	**88**	82	80	82	86	82	75	43	82	75	53	75	59	**96**	75	94	31	75	61	86	86	92	**96**	90	**96**	82	82	80	59	**96**	94	92	67	65	61	82	80	49	82	61	82	77
Ditrans. - cond. II simple	59	71	64	76	75	68	68	**86**	73	71	85	85	69	63	69	63	86	71	76	41	**92**	64	76	76	76	81	85	80	73	75	80	53	86	83	83	**93**	86	58	64	75	58	75	41	66	73
Ditrans. - future I progr.	61	52	51	**62**	61	57	**62**	57	51	90	28	**97**	90	11	90	80	57	62	67	39	85	56	48	56	74	67	72	89	59	62	69	11	69	57	66	70	36	52	49	57	43	**85**	54	61	61
Ditrans. - future I simple	88	60	51	56	**66**	56	61	**67**	45	64	60	73	64	58	64	70	72	67	91	43	64	76	58	59	72	57	62	**99**	65	61	61	18	62	59	59	67	53	49	55	43	47	**72**	44	58	61
Ditrans. - future II progr.	91	71	64	66	66	47	60	**86**	62	92	30	87	92	**99**	92	78	81	80	**100**	41	90	70	64	68	64	70	**99**	87	67	77	84	46	96	**100**	95	86	70	77	80	70	56	91	59	62	75
Ditrans. - future II simple	49	71	**94**	86	71	65	**94**	86	92	98	76	92	96	**100**	96	65	98	**100**	98	41	**100**	55	51	55	94	47	49	41	73	94	94	18	65	63	57	**98**	57	59	94	82	73	92	41	94	76
Ditrans. - past perf. progr.	91	60	44	62	**75**	66	58	**75**	48	75	68	85	75	33	75	75	87	79	79	40	**93**	33	79	81	80	63	74	78	76	64	65	59	56	82	66	**85**	77	76	56	58	47	58	54	58	67
Ditrans. - past perf. simple	112	63	62	65	63	**72**	71	65	61	60	66	57	58	36	58	58	70	**82**	64	47	66	69	41	44	54	55	45	67	62	64	57	39	54	57	59	65	58	52	**73**	57	39	67	52	70	59
Ditrans. - past progr.	83	58	59	71	52	59	61	**77**	60	48	61	51	46	22	46	43	**78**	36	**78**	37	61	71	40	35	39	**77**	76	**78**	51	63	60	20	63	69	**75**	67	**73**	31	59	52	51	53	53	66	56
Ditrans. - pres. perf. progr.	48	85	54	85	85	**92**	88	90	60	**100**	71	73	**100**	65	**100**	77	56	56	92	44	79	31	54	94	**100**	92	88	**100**	90	81	81	56	75	**96**	79	94	79	73	94	79	81	88	58	88	79
Ditrans. - pres. perf. simple	54	**65**	37	56	39	30	41	61	44	41	**93**	61	39	24	39	33	89	78	89	52	76	57	30	30	28	67	74	48	39	48	46	28	69	44	74	**87**	76	35	63	65	56	59	54	37	53
Ditrans. - pres. progr.	72	76	38	94	**97**	90	68	79	49	92	85	85	93	75	93	88	97	97	93	54	92	75	**100**	97	96	89	89	60	**97**	83	89	89	85	89	83	**96**	**97**	93	93	79	71	89	53	68	84
Ditrans. - simple past	77	77	56	77	79	56	66	**87**	57	92	82	95	92	**99**	92	73	**97**	92	**99**	45	96	53	61	68	92	92	87	82	81	84	81	45	68	88	71	**96**	91	71	78	60	60	79	42	64	77
Ditrans. - simple pres.	54	72	30	**83**	80	**83**	56	72	41	67	**98**	70	69	61	69	70	**98**	74	81	46	96	43	70	67	69	93	85	81	80	69	70	65	63	78	67	**98**	96	69	80	59	50	61	54	56	71
Gerund	161	93	90	**96**	**96**	86	83	93	86	**99**	**100**	**99**	**99**	84	**99**	90	93	**100**	94	41	98	13	96	94	95	98	97	60	96	89	91	**0**	**98**	**98**	94	**99**	**98**	95	92	77	90	94	66	86	87
Imperative	50	70	50	**96**	**96**	70	70	90	64	90	80	96	90	94	90	94	**100**	**100**	88	48	**100**	30	94	**100**	98	**100**	**100**	82	98	76	76	50	98	**100**	94	**100**	94	92	74	72	66	76	60	68	83
Intrans. - cond. I progr.	9	56	89	89	**100**	**100**	78	**100**	89	44	44	22	**100**	**100**	**100**	44	**100**	**100**	**100**	56	44	**100**	**100**	**100**	**100**	**100**	67	**100**	**100**	89	89	78	56	89	56	78	**100**	44	67	78	**100**	89	44	67	80
Intrans. - cond. I simple	3	**100**	**0**	67	**100**	**100**	33	**100**	33	**100**	33	33	**100**	**100**	**100**	33	**100**	**100**	**100**	33	33	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	67	67	67	**100**	67	67	**100**	33	**100**	**0**	67	67	**0**	33	73
Intrans. - future I progr.	7	71	86	**100**	86	57	**100**	86	86	14	**100**	86	71	**100**	71	57	**100**	**100**	**100**	43	**100**	86	71	71	71	**100**	**100**	86	86	**100**	**100**	43	**100**	86	86	**100**	**100**	86	**100**	**100**	43	57	29	**100**	82
Intrans. - future I simple	24	67	**75**	**75**	71	50	67	58	71	96	96	96	96	71	96	92	96	79	96	54	96	25	**100**	**100**	88	**100**	**100**	62	67	67	58	21	62	54	62	**96**	**96**	58	50	67	50	67	46	67	73
Intrans. - future II progr.	4	25	**50**	25	25	**50**	**50**	25	**50**	25	75	**100**	75	25	75	50	75	25	75	25	**100**	25	25	25	25	75	75	25	25	25	75	25	75	75	75	75	25	**100**	50	**100**	25	25	25	50	50
Intrans. - future II simple	7	71	**100**	86	**100**	**100**	**100**	57	**100**	**100**	43	86	**100**	71	**100**	43	86	**100**	71	14	86	**100**	**100**	86	**100**	71	71	29	**100**	86	**100**	14	86	71	57	86	43	43	**100**	71	43	**100**	29	**100**	77
Intrans. - past perf. progr.	16	56	50	38	62	69	62	**81**	69	44	44	25	62	56	62	**75**	44	56	38	56	38	62	38	50	56	50	50	**75**	69	62	**81**	6	31	75	44	38	44	50	38	56	50	62	44	62	53
Intrans. - past perf. simple	18	78	72	**89**	83	61	78	83	61	72	61	50	89	17	89	56	28	33	28	39	61	11	**100**	**100**	78	**100**	**100**	33	83	78	78	22	**100**	50	89	78	44	72	61	89	78	78	50	78	67
Intrans. - past progr.	28	43	57	**71**	68	54	57	68	54	64	46	64	64	68	64	61	71	**89**	39	25	61	18	46	50	50	68	86	68	**71**	54	50	39	64	54	64	61	61	64	57	46	50	**71**	50	57	58
Intrans. - pres. perf. simple	2	**100**	50	**100**	**100**	**100**	**100**	**100**	50	50	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**0**	**100**	50	50	50	50	**100**	**100**	**100**	**100**	**100**	**100**	50	**100**	**100**	50	**100**	**100**	**100**	**100**	**100**	**100**	**100**	50	**100**	86
Intrans. - pres. progr.	5	80	**100**	80	80	80	**100**	**100**	80	60	80	**100**	60	**100**	60	80	80	80	80	20	**100**	80	80	80	80	80	**100**	80	80	**100**	**100**	80	80	80	80	**100**	80	80	60	**100**	**100**	80	40	**100**	82
Intrans. - simple past	24	58	38	62	67	58	46	**79**	38	67	88	71	75	88	75	71	**100**	**100**	**100**	58	71	38	67	67	62	71	96	**100**	67	71	67	42	71	**92**	75	75	75	75	62	54	50	67	54	46	69
Intrans. - simple pres.	10	40	30	50	50	50	40	**80**	40	70	60	50	90	90	90	70	80	80	70	50	90	70	**100**	90	70	40	80	80	50	30	40	60	60	70	50	**80**	**80**	50	40	20	40	50	70	40	61
Modal	20	**70**	60	40	50	45	55	50	55	35	60	45	35	**0**	35	**100**	10	5	5	45	30	75	90	85	95	15	5	55	50	**75**	70	15	30	30	30	40	30	10	50	55	35	40	70	60	45
Modal negated	20	35	65	70	75	60	60	**95**	70	85	95	85	85	50	80	60	90	**100**	**100**	20	85	35	55	50	75	85	70	85	80	70	75	45	80	**85**	75	**85**	**85**	**85**	55	45	45	75	60	60	70
Reflex. - cond. I progr.	65	66	52	48	45	45	46	34	**71**	20	40	69	37	**95**	37	46	58	**95**	77	42	82	29	34	40	49	35	75	12	45	54	66	43	77	82	**89**	66	55	82	45	71	46	46	51	51	55
Reflex. - cond. I simple	112	**76**	70	48	59	58	70	32	72	48	71	60	49	84	49	43	92	94	**97**	41	70	78	49	60	70	37	88	5	59	66	82	72	89	71	93	69	74	80	68	**96**	78	62	69	69	67
Reflex. - cond. II progr.	97	**71**	**72**	66	66	61	69	49	**71**	63	60	66	75	92	75	28	81	**96**	89	51	89	36	47	51	74	65	82	49	66	67	81	65	80	**87**	81	81	64	79	72	**87**	58	58	72	72	69
Reflex. - cond. II simple	109	61	**68**	52	59	54	61	52	59	74	56	61	81	81	81	16	86	85	86	46	**90**	52	28	38	76	56	87	55	57	47	68	67	87	78	89	**91**	72	77	53	**92**	41	64	63	63	65
Reflex. - future I progr.	70	67	67	70	57	**84**	79	41	66	56	69	71	56	79	56	69	59	**97**	81	51	81	60	51	47	66	74	79	29	60	66	74	46	77	77	69	77	63	76	61	**83**	53	71	77	77	67
Reflex. - future I simple	83	70	67	73	58	82	**86**	58	78	52	49	34	57	70	57	78	47	**86**	59	31	60	57	69	58	77	58	60	55	59	76	80	82	72	67	77	64	77	61	63	**87**	51	76	**87**	**87**	66

Accuracy of the metrics(%) with regards to the linguistically-motivated phenomena for English-German. Continued on next page

ling. phenomenon	#	baselines								QE as a metric																			ref. based metrics																	avg
		BERTscore	BLEU	BLEURT-20	COMET	YiSi-1	chrF	prismRef	spBLEU	Calibri-COMET22-QE	CometKiwi-XL	CometKiwi-XXL	CometKiwi	GEMBA-MQM	KG-BERT	MS-COMET-QE	MetricX-23-QE-b	MetricX-23-QE-c	MetricX-23-QE	Random-sysname	XCOMET-QE-Ensemble	XLsimQE	cometoid22-wmt21	cometoid22-wmt22	cometoid22-wmt23	mltsamp-bleurt0p2p1-qe	mltsamp-bleurtxv1p-qe	prismSrc2	Calibri-COMET22	MEE4	MEE4_stsb_xlm	MaTESe	MetricX-23-b	MetricX-23-c	MetricX-23	XCOMET-Ensemble	XCOMET-XL	XCOMET-XXL	XLsim	eBLEU	embed_llama	mt0regressor	partokengram_F	tokengram_F	
Reflex. - future II progr.	81	65	56	64	64	75	**80**	68	57	68	75	58	85	77	85	52	85	**89**	86	51	83	37	44	49	72	64	85	41	67	72	79	57	84	73	80	**86**	81	81	68	79	38	68	77	77	69
Reflex. - future II simple	56	71	66	77	54	**88**	**88**	71	64	77	95	89	84	73	84	39	98	98	**100**	50	**100**	70	66	64	68	48	54	38	52	75	82	61	89	54	89	**98**	68	64	64	88	43	64	89	89	73
Reflex. - past perf. progr.	98	60	50	67	64	**71**	66	49	51	81	79	70	**88**	85	**88**	52	74	73	83	39	77	48	48	61	83	67	77	43	65	67	69	41	66	71	70	**81**	70	73	61	72	49	73	65	65	66
Reflex. - past perf. simple	53	62	47	68	55	**74**	55	49	57	70	89	85	77	64	77	17	81	79	79	38	**98**	60	45	51	68	70	83	49	58	62	58	49	81	70	91	**92**	83	72	57	77	47	83	53	53	66
Reflex. - past progr.	5	**100**	**100**	40	**100**	**100**	**100**	80	**100**	**0**	60	60	40	40	40	**100**	**100**	**100**	**100**	40	40	40	40	60	40	40	60	40	**100**	**100**	**100**	60	80	60	80	60	80	60	**100**	**100**	**100**	40	**100**	**100**	72
Reflex. - pres. perf. progr.	33	76	48	**88**	67	76	76	45	48	61	**100**	88	88	85	88	48	**100**	**100**	**100**	42	**100**	27	61	79	82	67	**100**	24	67	79	85	52	97	79	**100**	**100**	97	**100**	67	**100**	33	73	79	79	76
Reflex. - pres. perf. simple	39	59	46	67	59	69	**72**	46	44	44	**97**	72	69	64	69	31	74	85	85	38	82	62	38	67	59	56	74	31	59	69	79	28	79	69	85	74	77	82	67	**90**	54	69	69	69	65
Reflex. - pres. progr.	99	62	51	54	56	**67**	56	37	62	51	75	66	46	**91**	46	39	58	76	70	41	79	51	58	56	52	57	44	45	58	57	64	51	71	69	74	73	**76**	70	59	72	42	61	58	58	59
Reflex. - simple past	119	71	70	73	73	73	**77**	64	71	63	73	81	82	82	82	46	95	98	**100**	34	86	55	55	60	75	61	54	63	72	71	77	57	73	**84**	71	**85**	82	77	74	79	56	82	78	78	72
Reflex. - simple pres.	138	65	65	67	67	**88**	63	64	67	55	70	75	58	87	58	47	78	**97**	73	49	71	31	56	57	72	78	54	64	67	69	80	43	70	67	70	77	72	67	77	**83**	54	67	64	64	67
trans. - future II progr.	11	73	82	73	**91**	64	82	82	82	**100**	73	**100**	**100**	91	**100**	91	**100**	**100**	**100**	36	91	73	91	82	91	**100**	**100**	64	91	73	91	36	**100**	**100**	**100**	91	82	91	73	82	73	73	82	82	84
trans. - cond. I progr.	11	55	**91**	45	64	36	82	55	**91**	36	91	**100**	55	36	55	82	91	45	**100**	36	91	73	73	73	64	**100**	**100**	55	55	55	55	55	82	73	73	**100**	**100**	82	55	45	55	18	82	82	68
trans. - cond. I simple	9	67	**100**	89	89	56	**100**	22	**100**	67	89	**100**	56	56	56	89	**100**	67	89	33	89	89	44	33	89	**100**	**100**	**0**	67	67	56	56	**100**	67	**100**	**100**	**100**	78	67	67	67	56	**100**	**100**	75
trans. - cond. II progr.	20	70	55	75	65	80	55	**85**	60	70	**95**	**95**	70	75	70	40	75	85	90	40	85	40	65	65	60	85	**95**	60	65	75	70	50	70	**90**	70	**90**	85	85	65	55	55	75	55	55	70
trans. - cond. II simple	2	50	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	50	**100**	**100**	**100**	**100**	**100**	**100**	**100**	50	50	50	**100**	**100**	**100**	50	**100**	**100**	**100**	50	**100**	**100**	**100**	**100**	**100**	**100**	**100**	50	50	**100**	**100**	**100**	90
trans. - future I progr.	12	42	**83**	75	**83**	25	50	75	75	75	75	75	75	58	75	58	**92**	67	**92**	58	75	67	75	75	75	**92**	**92**	58	75	50	50	58	67	**92**	67	**92**	**92**	**92**	50	42	42	42	67	67	69
trans. - future I simple	22	68	**95**	64	68	68	77	50	**95**	73	86	36	77	5	77	86	95	73	**100**	50	59	73	77	73	68	64	91	59	64	68	68	73	86	50	86	77	**91**	73	45	68	59	41	86	86	71
trans. - future II simple	39	62	**92**	59	77	67	85	62	90	85	72	79	87	54	87	69	74	69	72	38	**97**	72	49	44	87	67	62	33	77	82	79	33	67	62	67	79	69	69	74	69	59	74	**87**	**87**	70
trans. - past perf. progr.	16	50	69	50	**81**	**81**	**81**	56	69	**94**	**94**	38	**94**	31	**94**	75	81	50	81	38	50	62	75	69	62	44	75	62	**81**	56	56	56	69	50	75	75	**81**	38	75	62	50	44	75	75	66
trans. - past perf. simple	9	44	78	**89**	**89**	33	**89**	22	78	**100**	89	**100**	**100**	44	**100**	56	**100**	78	89	22	**100**	78	89	67	67	**100**	**100**	**0**	**89**	44	44	56	78	**89**	78	**89**	**89**	78	33	33	22	22	78	78	70
trans. - pres. perf. progr.	5	20	**80**	**80**	**80**	20	**80**	20	**80**	80	**100**	**100**	**100**	**0**	**100**	**100**	**100**	20	80	60	**100**	20	**100**	**100**	**100**	**100**	**100**	40	80	20	**0**	80	60	**100**	60	**100**	80	80	20	20	20	**0**	80	80	66
trans. - pres. perf. simple	9	33	67	**78**	**78**	44	**78**	44	67	**100**	89	**100**	**100**	22	**100**	**100**	**100**	44	**100**	33	89	67	78	78	78	**100**	**100**	44	78	33	22	56	78	**100**	78	89	89	89	22	22	33	33	78	78	70
trans. - pres. progr.	10	30	**70**	20	40	30	40	30	50	30	40	50	30	20	30	40	**90**	60	**90**	50	60	50	50	50	30	40	40	40	50	40	30	20	40	60	40	**70**	**70**	60	40	40	20	40	50	50	45
trans. - simple past	23	61	43	**96**	70	35	57	87	52	91	78	91	**96**	61	**96**	91	**96**	74	83	48	**96**	39	74	74	87	**96**	**96**	78	70	65	70	52	96	**100**	91	96	91	**100**	57	48	43	70	57	57	75
trans. - simple pres.	16	31	62	38	**75**	69	62	56	56	94	50	81	**100**	62	**100**	**100**	94	94	75	50	81	50	**100**	94	**100**	81	81	69	62	38	50	**81**	75	75	56	**81**	69	75	56	56	44	75	62	62	70
Verb valency																																													
Case government	57	82	67	75	82	**86**	70	**86**	75	75	74	81	75	75	75	72	86	**91**	82	40	82	40	81	79	84	84	89	65	75	77	75	42	**82**	**82**	79	81	75	79	81	65	79	75	72	72	76
Catenative verb	177	71	62	**87**	75	68	63	68	63	73	91	**94**	73	34	73	66	88	63	90	44	92	60	77	81	67	92	89	58	72	65	68	19	81	71	84	**90**	85	77	80	64	64	63	64	64	72
Middle voice	29	83	62	**97**	93	72	76	93	72	93	79	97	93	86	93	62	**100**	**100**	**100**	55	**100**	21	**100**	93	**100**	97	**100**	31	90	90	83	59	**100**	**100**	**100**	**100**	**100**	90	79	59	83	83	76	76	84
Passive voice	70	66	54	69	70	67	74	**81**	61	74	71	63	79	94	79	69	96	**99**	90	40	79	30	67	63	70	74	89	87	70	64	69	33	**89**	81	87	71	59	51	80	57	69	60	71	71	71
Resultative	147	73	69	84	83	81	75	**86**	76	79	76	84	78	71	78	80	78	84	77	37	83	65	82	84	84	**88**	84	73	83	73	69	51	84	**88**	86	82	80	80	75	75	67	61	75	75	77
macro avg.	8945	67	65	74	**75**	70	71	73	67	73	76	77	78	67	78	68	**84**	80	**83**	43	82	52	71	72	77	78	**82**	62	74	70	72	42	79	80	77	**84**	79	71	70	67	58	69	59	71	71
micro avg.	8945	70	65	**75**	**76**	72	69	74	68	73	74	75	74	68	74	67	79	**81**	80	44	78	47	71	71	78	77	**81**	62	74	71	73	36	79	**81**	78	**81**	77	70	73	69	59	69	56	70	71

Table 5: Accuracy of the metrics(%) with regards to the linguistically-motivated phenomena for English-German

			baselines								QE as a metric																ref. based metrics											
ling. category	ling. phenomenon	#	BERTscore	BLEU	BLEURT-20	COMET	YiSi-1	chrF	prismRef	spBLEU	CometKiwi-XL	CometKiwi-XXL	CometKiwi	GEMBA-MQM	KG-BERT	MS-COMET-QE	MetricX-23-QE-b	MetricX-23-QE-c	MetricX-23-QE	Random-sysname	XCOMET-QE-Ensemble	XLsimQE	cometoid22-wmt21	cometoid22-wmt22	cometoid22-wmt23	prismSrc2	MaTESe	MetricX-23-b	MetricX-23-c	MetricX-23	XCOMET-Ensemble	XCOMET-XL	XCOMET-XXL	XLsim	eBLEU	embed_llama	partokengram_F	tokengram_F	avg
Ambiguity	Lexical ambiguity	66	80	67	**100**	**97**	88	67	83	73	**100**	94	**100**	**100**	**100**	95	**100**	**100**	**100**	39	**100**	30	**100**	**100**	**100**	56	**98**	**100**	**100**	**100**	**100**	**100**	**100**	88	77	44	59	70	86
Coordination & ellipsis	Gapping	44	**80**	75	68	**77**	73	66	**77**	73	93	86	66	30	66	93	43	**98**	41	52	84	34	73	68	80	68	14	45	75	39	73	70	57	73	**86**	50	55	75	66
	Pseudogapping	45	87	78	**89**	**89**	**91**	87	**91**	80	**87**	80	76	40	76	69	69	69	56	47	71	53	78	67	78	53	13	64	73	49	73	60	44	**89**	84	78	62	80	70
	Sluicing	11	64	55	**82**	45	55	64	55	55	18	55	36	64	36	64	**82**	**82**	**82**	45	64	73	55	55	27	27	0	55	27	36	64	55	0	**91**	**91**	64	45	64	54
	Stripping	27	78	63	59	48	63	56	**81**	52	41	26	26	15	26	52	67	22	22	37	22	48	52	56	48	**74**	4	30	37	37	41	67	41	67	67	**85**	22	52	47
	VP-ellipsis	76	**91**	76	57	74	76	59	70	68	72	57	86	47	86	82	63	**93**	**95**	43	62	32	74	74	71	46	39	**79**	**82**	**79**	68	62	61	66	76	45	70	61	68
False friends	False friends	6	**100**	17	**100**	**100**	**100**	17	83	50	0	0	0	**100**	0	0	0	**100**	17	17	0	17	33	33	**100**	0	17	**100**	**100**	**100**	33	33	**100**	83	**100**	50	83	33	50
Function word	Focus particle	50	**56**	44	36	30	52	36	**56**	46	42	24	42	20	42	18	48	44	**62**	44	56	34	48	26	10	46	16	66	**70**	64	48	36	20	42	42	52	56	38	42
	Question tag	66	62	88	**94**	82	53	39	42	**91**	62	79	56	74	56	50	80	64	38	44	85	42	**92**	**92**	88	14	3	61	47	58	70	**79**	24	55	**76**	52	52	38	61
MWE	Collocation	79	71	68	**81**	**82**	76	67	73	72	86	**87**	**90**	65	**90**	**90**	82	86	85	39	85	34	85	85	85	34	46	82	**86**	84	**89**	84	81	70	72	70	65	66	75
	Idiom	31	61	45	65	68	58	42	**84**	42	26	13	39	39	39	68	55	**90**	45	48	10	16	52	32	52	0	90	**100**	**100**	90	13	35	32	81	45	39	45	45	50
	Verbal MWE	12	83	83	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	92	50	92	17	83	92	83	25	**100**	92	75	67	92	8	42	83	**100**	83	**100**	**100**	83	**100**	83	42	58	**100**	81
Named entity & terminology	Date	45	60	47	**87**	69	58	49	80	49	80	78	53	40	53	64	69	56	**87**	42	76	38	82	82	80	58	9	84	67	84	82	**100**	53	73	64	67	53	49	65
	Domainspecific term	150	82	71	**97**	**99**	85	79	93	81	93	57	53	65	53	92	61	87	77	51	61	44	81	83	**98**	8	68	**97**	**100**	**97**	95	**99**	94	85	81	65	51	79	77
	Measuring unit	45	87	84	**100**	84	82	89	62	84	53	60	82	33	82	2	69	42	42	47	60	**100**	**100**	**98**	**100**	80	33	64	60	47	82	53	11	**91**	62	62	87	87	68
	Proper name	3	**100**	67	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	0	**100**	0	0	**100**	**100**	67	**100**	33	**100**	**100**	**100**	33	67	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	85
Negation	Negation	34	74	74	79	68	74	**85**	68	74	68	85	85	91	85	41	71	**100**	9	59	91	9	62	29	59	91	0	**94**	71	**94**	79	71	74	65	76	76	85	85	69
Non-verbal agreement	Coreference	57	68	54	**86**	77	67	67	60	56	84	**100**	56	77	56	75	75	**100**	67	42	75	26	**100**	**100**	84	63	46	**93**	**93**	**93**	84	82	33	61	60	63	68	68	71
	Genitive	4	**75**	**75**	25	**75**	**75**	**75**	**75**	**75**	75	**100**	25	0	25	50	25	75	50	25	**100**	25	25	**100**	**100**	50	0	75	75	75	**100**	25	0	**100**	75	**100**	75	75	60
Punctuation	Direct Speech	121	42	46	53	68	**86**	41	73	55	66	29	65	27	65	**87**	55	66	43	41	61	44	66	67	36	56	1	60	71	52	66	57	12	63	66	**84**	45	45	54
Subordination	Adverbial clause	132	43	52	32	47	**55**	46	48	47	49	**74**	41	32	41	70	55	55	22	49	58	16	48	16	39	22	5	60	**92**	44	64	43	80	39	30	39	45	45	46
	Cleft sentence	24	71	54	**92**	**92**	**92**	75	79	54	96	79	75	88	75	79	92	96	**100**	54	**100**	46	88	88	88	67	79	88	92	92	**100**	92	92	75	54	75	62	62	80
	Contact clause	10	60	30	90	**100**	70	60	**100**	30	60	**100**	**100**	0	**100**	**100**	40	**100**	**100**	70	**100**	**100**	**100**	**100**	**100**	**100**	0	**100**	**100**	**100**	**100**	60	0	80	40	60	60	60	74
	Infinitive clause	24	71	75	67	**96**	79	**96**	67	67	**96**	54	83	62	83	83	71	88	75	38	83	38	75	75	79	79	33	92	**100**	92	96	96	50	71	71	58	96	96	76
	Object clause	70	66	80	73	**83**	67	**84**	79	**83**	**100**	93	**100**	23	**100**	84	21	19	21	34	**100**	37	**100**	**100**	**100**	50	0	26	80	26	**100**	84	17	71	71	57	89	89	67
	Participle clause	33	70	82	67	79	**94**	76	42	82	42	70	42	9	42	9	**100**	52	94	45	58	12	9	9	30	9	0	**82**	70	**79**	52	33	55	76	70	61	76	76	54
	Pseudo-cleft sentence	51	**84**	78	**86**	**86**	78	65	73	73	65	63	65	69	65	61	75	**100**	88	39	75	55	94	84	88	43	35	75	78	76	76	78	63	**92**	71	65	63	63	72
	Relative clause	117	66	56	66	**71**	63	**74**	66	67	81	41	64	35	64	**91**	85	78	86	47	83	36	62	74	66	25	32	54	**85**	68	**84**	**85**	32	67	58	60	71	71	64
	Subject clause	38	55	53	53	55	**61**	45	**61**	47	84	58	**97**	82	**97**	84	61	76	84	47	84	24	84	82	53	63	5	50	68	37	**84**	**87**	47	53	45	79	50	50	62
Verb tense/aspect/mood	Conditional	6	**100**	67	**100**	**100**	**100**	83	**100**	83	0	17	**83**	17	**83**	0	67	17	33	17	17	67	0	0	**83**	50	33	**100**	**100**	**100**	67	67	50	**100**	50	**100**	83	83	62
	Ditransitive	13	85	69	77	**100**	69	92	**100**	69	69	**92**	**92**	62	**92**	69	69	62	**92**	54	69	46	**92**	77	**92**	31	85	**100**	**100**	**100**	92	92	85	62	69	62	85	85	79
	Gerund	34	82	74	**100**	**100**	88	82	91	85	91	**97**	91	88	91	85	91	91	94	38	**100**	26	**100**	**100**	**100**	18	74	91	**97**	91	91	82	79	79	71	68	82	82	83
	Imperative	12	**75**	67	**75**	**75**	58	**75**	**75**	**75**	**100**	**100**	83	83	83	75	**100**	**100**	92	33	**100**	75	75	50	83	33	17	**100**	**100**	**100**	**100**	**100**	**100**	92	75	92	75	75	80
	Intransitive	3	67	33	67	67	67	**100**	67	67	**100**	**100**	67	**100**	67	**100**	**100**	**100**	**100**	0	**100**	**100**	67	67	**100**	0	0	**100**	**100**	**100**	**100**	**100**	**100**	67	**100**	33	67	67	76
	Reflexive	24	71	54	62	62	58	**88**	71	75	**100**	92	71	33	71	83	71	79	71	50	67	88	88	58	83	50	17	**100**	92	**100**	96	**100**	75	83	67	83	83	83	74
	Transitive	43	77	56	63	**84**	79	**81**	72	65	65	53	49	51	49	37	35	33	28	35	58	67	37	33	**72**	51	53	77	88	77	70	**95**	60	74	70	67	84	84	62
Verb valency	Case government	3	**100**	33	0	33	**100**	33	33	33	**67**	**67**	**67**	0	**67**	0	0	0	33	0	**67**	33	0	0	0	33	33	0	**67**	33	**67**	**67**	0	33	**67**	33	33	33	35
	Catenative verb	50	78	80	74	**90**	86	70	66	80	80	**94**	**96**	56	**96**	90	74	90	74	36	78	42	72	66	76	42	58	76	82	74	74	70	66	90	**92**	**94**	74	74	75
	Impersonal Subject	10	40	50	**80**	40	70	50	20	40	40	**100**	40	90	40	40	**100**	**100**	**100**	40	40	0	40	40	40	10	60	**100**	**100**	**100**	40	40	40	**100**	40	40	50	50	56
	Resultative	58	67	60	**69**	**71**	67	**69**	55	64	69	**83**	71	74	71	64	78	**83**	76	36	**81**	17	62	64	64	12	31	**81**	78	**79**	76	76	72	**79**	69	45	69	69	65
macro avg.		1727	73	62	74	**77**	**75**	68	72	67	70	71	68	51	68	60	65	**75**	67	40	**72**	44	68	65	**73**	41	31	77	**83**	76	77	73	55	76	69	64	66	68	66
micro avg.		1727	69	64	72	**76**	73	65	70	67	**74**	67	68	51	68	**71**	67	**74**	64	44	**72**	39	**73**	68	**71**	40	32	73	**82**	71	77	74	55	71	67	62	63	66	66

Table 6: Accuracy of the metrics(%) with regards to the linguistically-motivated phenomena for English-Russian

Tokengram_F, a fast and accurate token-based chrF++ derivative

Sören Dréano
ML-Labs
Dublin City University
soren.dreano2@mail.dcu.ie

Derek Molloy
School of Electronic Engineering
Dublin City University
derek.molloy@dcu.ie

Noel Murphy
School of Electronic Engineering
Dublin City University
noel.murphy@dcu.ie

Abstract

The tokengram_F metric presented in this paper is a novel approach to evaluating machine translation that has been submitted as part of the WMT23 challenge. It offers a new perspective on evaluating machine translation that takes advantage of modern tokenization algorithms to provide a more natural representation of the language in comparison to word *n*-grams.

Tokengram_F is an F-score-based evaluation metric for Machine Translation that is heavily inspired by chrF++ and can act as a more accurate replacement. By replacing word *n*-grams with *n*-grams obtained from tokenization algorithms, tokengram_F captures similarities between words sharing the same semantic roots.

While requiring minimal training based on an open corpus of monolingual datasets, the tokengram_F metric proposed still retains excellent performance that is comparable to more computationally expensive metrics. The tokengram_F metric demonstrates its versatility by showing satisfactory results, even when a tokenizer for a specific language is not available. In such cases, the tokenizer of a related language can be used instead, highlighting the adaptability of the tokengram_F metric to less commonly-used languages.

1 Introduction

Machine Translation (MT) is a subdomain of Neural Language Processing (NLP) that is focused on the translation of one natural language to another, with the aim of producing natural-sounding sentences. To evaluate the quality of algorithm-generated translations, evaluation metrics provide quantitative scores to objectively assess the accuracy of the model. Machine-generated translations are compared to human-generated translations in different ways depending on the evaluation metric. In recent years, machine translation has seen a great deal of progress in terms of accuracy and fluency. However, there is still a need for more robust evaluation metrics that can effectively measure the quality of machine-generated translations.

Popular metrics in MT include BLEU (Papineni et al., 2002), which measures the overlap between sequences of words in the reference and generated texts; chrF (Popović, 2015), which is an F1 score at the character-level; chrF++ (Popović, 2017), which extends chrF with word *n*-grams; TER (Snover et al., 2006), which counts the erroneously-aligned words between the reference and the generation, and METEOR (Banerjee and Lavie, 2005), which takes into account synonyms and stems.

More recent metrics rely on neural network architectures, such as COMET (Rei et al., 2020) and MS-COMET (Kocmi et al., 2022). By using the similarity of vector representations of the generated translation and the reference translation, they provide state-of-the-art machine evaluation of translations at the cost of being expensive to train and compute.

Since its third instance in 2006, the Workshop on Statistical Machine Translation (WMT) has released an evaluation task to compare metrics each year. Generated translations are usually ranked by humans, and the correlation coefficient between the human-performed ranking and the evaluation metric-performed ranking determines the quality of the metric.

2 Tokengram_F

2.1 Tokenization

As text cannot be directly processed by machine learning algorithms, it first has to be converted into a numerical representation. Tokenization splits the text into smaller character sequences, including but not limited to phonemes, syllables, letters, words or base pairs, collectively named tokens. The tokenization process also often consists of adding special tokens, such as the unknown <unk> token to represent never-seen characters or the padding

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 730–737
December 6–7, 2023.

<pad> token to pad the sentence to a fixed length. Each token can be converted to and from a unique identifier, which is usually an integer between 0 and the maximum vocabulary size minus one.

2.2 chrF++

An n-gram refers to a consecutive series of n tokens that are extracted from a given corpus of text or speech, with these units of text being defined based on the particular context of the application. A character-gram, or unigram, is a token that contains exactly one character, while a word-gram contains an entire word. chrF++ is an F-score using both word-grams and character-grams to compare the generated translation to the reference translation. The general formula is

$$ngrF\beta = (1 + \beta^2)\frac{(ngrP \times ngrF)}{(\beta^2 \times ngrP + ngrR)}$$

where β determines the weight of the recall as discussed in Section 3.4.3.

2.3 Modern tokenization algorithms

Subword-based tokenization divides words depending on their number of occurrences in the training data. Subwords can be combined to represent less frequent words or even words that were not present in the training data. For instance, in cases where a word such as "decaying" is absent from the vocabulary, an English tokenizer may represent it by combining the "decay" and "ing" tokens.

Byte Pair Encoding (BPE) (Sennrich et al., 2016), had been used for data compression long before it was ever applied in NLP-related tasks. After first counting the frequency of each unique word in the training data, BPE merges frequent occurrences of subword pairs until it reaches the desired vocabulary size.

Instead of starting from a small vocabulary representing the set of unique words, and growing in size from there (as in BPE), Unigram (Kudo, 2018) initialises its base vocabulary to a large number of symbols and then trims it down to the desired size. It is analogous to factor analysis, as at each step it calculates the loss of information that would be induced by removing each token, and then erases the less important ones from its vocabulary.

The sentence "The kingly sovereign governs" becomes:

```
words: "The" "kingly" "sovereign" "governs"
tokens: "The" "king" "ly" "sovereign" "govern" "s"
```

2.4 Replacing word-grams

When using word-grams for scoring, each word is compared regardless of its proximity to other words. For example, the words "say" and "saying" share a common root but this link would be lost when using word-grams.

In this work, the authors claim that modern tokenization algorithms can be used instead of word-grams to split the text in a more natural manner that reflects the structure of each language. Tokengram_F is an evaluation metric derived from chrF++ that replaces the use of word-grams by tokens learned either by Unigram or by BPE.

3 Methodology

3.1 Framework

SentencePiece (Kudo and Richardson, 2018) is a fast data-driven text tokenizer and detokenizer implementing the Unigram algorithm. The vocabulary size (number of individual tokens) needs to be provided before training. The minimum vocabulary size would consist of the number of special tokens and individual characters of the alphabet for each language. A large vocabulary size might lead to overfitting and a reduction in the effectiveness of the model, given that some parameters will be dedicated to rare words.

3.2 Training

The tokengram_F score uses the same 3-letter ISO-639-2 language code as the Tatoeba dataset, while the WMT tasks rely on the 2-letter ISO-639-1 language code. The website of the Library of Congress (Library of Congress, 2017) was used for conversions between the two norms.

The Tatoeba Translation Challenge (Tiedemann, 2020) is an initiative that aims to evaluate the effectiveness of MT systems on a large, diverse, and high-quality parallel corpus. While the main training data relies on OPUS (Tiedemann, 2012), which provides open-source sentence-aligned text corpora to support data-driven NLP, Tatoeba also provides monolingual datasets extracted from CirrusSearch Wikimedia dumps (Foundation, 2023).

Out of the 279 different languages available, 240 had a sufficiently large corpus to be included in the work described in this paper.

3.3 Exceptions

As the tokengram_F metric is dependent on the utilization of the Tatoeba monolingual datasets for tokenizer training, adaptations were necessary to accommodate languages that are not represented within this dataset.

3.3.1 Livonian:

While there is no dataset in the Tatoeba Translation Challenge to train a Livonian tokenizer, the Latvian tokenizer produced satisfactory results and was utilised as a substitute, highlighting the versatility of the tokengram_F metric and its ability to accommodate languages that are less frequently used.

3.3.2 Serbian and Indonesian:

The Tatoeba challenge does not offer monolingual datasets for neither Serbian nor Indonesian. Nevertheless, the Tatoeba Wikimedia data, which are appropriate for tokenizer training and available for both the Serbian and Indonesian languages, were employed as a substitute.

3.4 Tokengram_F parameters

3.4.1 Tokenization algorithm:

While Tokengram_F can be used with any tokenization algorithm, this study examined both BPE and Unigram.

3.4.2 *n*-gram length:

This parameter determines the number of items in the reference that will be compared with each item in the source sentence, to assess the degree of correspondence of the two sentences.

Previous work (Popović, 2015, 2017) has indicated that for chrF++ there is no necessity to set the maximum word *n*-gram length beyond N=6.

3.4.3 Beta:

In this metric, the relative importance of precision and recall in the evaluation metric is determined by the β parameter. When beta is equal to 1.0, precision and recall have equal importance, while when beta is equal to 3.0, recall is three times more significant than precision. Previous research (Popović, 2015) has evaluated two beta values, 1.0 and 3.0, with the latter being considered "the most promising variant" due to the higher correlations it obtained.

3.4.4 Vocabulary size:

The goal of the present study is to mitigate the effect of infrequent words on the accuracy of the tokengram_F metric. To investigate the influence of vocabulary size on the performance of the metric, three tokenizers were trained for each language using vocabulary sizes of 16,000, 32,000, and 50,000 tokens. As the average vocabulary size tends to decrease from one year to the next (Libovický, 2021), wider vocabularies have not been examined.

3.5 Optimal parameters

3.5.1 Finetuning

The optimal parameters were determined based on achieving the highest average correlation among segments or systems across three datasets: WMT20 (Mathur et al., 2020), WMT21 (Freitag et al., 2021), and WMT22 (Freitag et al., 2022).

Initially, the *n*-gram length was assessed at values of 3, 6, and 9, while maintaining a vocabulary size of 50,000. Consistent with the findings of the original paper, a *n*-gram length of 6 demonstrated the strongest correlation as shown in Table 1, and thus it was chosen for subsequent evaluations.

Subsequently, the beta values of 2, 3, and 4 were examined specifically for an *n*-gram length N of 6. The best overall correlation is obtained with Unigram and β=3.0.

Table 3 presents the results obtained with a vocabulary size of 32,000. As with the previous vocabulary size, the choice of the tokenization algorithm only slightly affects the results. A β of 3.0 or 4.0 seems to give the best results.

As shown in Table 2, the vocabulary size of 16,000, which was the smallest size examined, exhibits generally weaker correlations compared to larger sizes, thus precluding the exploration of smaller sizes.

3.5.2 Results

Despite the marginal disparity, when the results are not rounded, the optimal parameters for tokengram_F are found to be a vocabulary size of 50,000, unigram tokenization, an *n*-gram length N=6, and a β value of 3.0.

3.6 Source code

The source code of tokengram_F is available at https://github.com/SorenDreano/tokengram_F.

4 Conclusion

Instead of using word *n*-grams as a basis for comparing a generated translation with a reference translation, tokengram_F utilises contemporary tokenization algorithms to accomplish this task. As a result, words that share common roots are deemed similar, regardless of whether they are exact matches or not.

The results obtained from evaluating the tokengram_F metric on the WMT20, WMT21, and WMT22 datasets indicate that the use of tokens generated through the SentencePiece framework leads to improved performance compared to the use of traditional word-grams in the chrF++ metric. Full results are displayed in Appendix A. The tokengram_F metric is a simple and efficient method for obtaining a reasonable correlation with human rankings, with the added benefit of requiring minimal training time to be applied to new languages.

In the segment-level task of the WMT22 edition, tokengram_F managed to obtain better overall correlations than any other metric that could provide results for all language pairs, including ones that require extensive neural networks to operate. With the exception of two tasks, tokengram_F outperformed both chrF and chrF++ metrics.

In conclusion, the tokengram_F metric is presented as a promising alternative for evaluating the quality of machine translations, as it offers a simple and efficient solution with above-average performance compared to other models. The findings of this study provide strong evidence of the potential of the tokengram_F metric as a valuable evaluation tool for machine translation. Its combination of simplicity, efficiency, and adaptability make it an attractive alternative to existing metrics and a promising direction for future research in the field.

5 Further work

The optimal number of tokens in a tokenizer may vary depending on the language. Subsequent research could concentrate on determining the most suitable vocabulary size per language.

The majority of the tokenizers were trained using the MonoLinguage Datasets from the Tatoeba Challenge, which are based on data from the Wikimedia Foundation. It remains possible that alternate data sources may produce varying results.

6 Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Wikimedia Foundation. 2023. Wikimedia downloads.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André Martins. 2022. Results of wmt22 metrics shared task: Stop using bleu - neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation*, pages 46–68, Abu Dhabi.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Tom Kocmi, Hitokazu Matsushita, and Christian Federmann. 2022. MS-COMET: More and better human judgements improve metric performance. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 541–548, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. *CoRR*, abs/1804.10959.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Jindřich Libovický. 2021. Jindřich's blog – machine translation weekly 86: The wisdom of the wmt crowd. Online, Accessed: 07.02. 2023.

Library of Congress. 2017. Codes for the representation of names of languages.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. *CoRR*, abs/2009.09025.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörg Tiedemann. 2020. The tatoeba translation challenge - realistic data sets for low resource and multilingual MT. *CoRR*, abs/2010.06354.

A Appendix. Tables

Table 1: Correlations over all metrics depending of the hyperparameters with a vocabulary size of 50,000

Vocabulary	50000									
Algorithm	BPE					Unigram				
Beta	4	3			2	4	3			2
n-gram	6	9	6	3	6	6	9	6	3	6
System WMT20	0.875	0.871	0.876	0.871	0.877	0.876	0.871	0.876	0.871	0.877
System WMT21	0.715	0.715	0.716	0.713	0.717	0.716	0.715	0.717	0.714	0.718
System WMT22	0.837	0.831	0.834	0.834	0.829	0.836	0.831	0.835	0.834	0.830
Segment WMT20	0.277	0.274	0.277	0.279	0.277	0.276	0.273	0.277	0.278	0.277
Segment WMT21	0.158	0.155	0.158	0.166	0.157	0.160	0.159	0.158	0.170	0.152
Segment WMT22	0.398	0.393	0.398	0.399	0.400	0.398	0.393	0.399	0.397	0.400
Average	0.543	0.540	**0.544**	0.543	0.543	**0.544**	0.540	**0.544**	0.543	0.542

Table 2: Correlations over all metrics depending of the hyperparameters with a vocabulary size of 16,000

Vocabulary size	16000					
Tokenization algorithm	BPE			Unigram		
Beta	4	3	2	4	3	2
n-gram length	6					
System WMT20	0.875	0.876	0.877	0.876	0.877	0.877
System WMT21	0.715	0.716	0.717	0.716	0.717	0.718
System WMT22	0.839	0.837	0.83	0.839	0.837	0.83
Segment WMT20	0.278	0.278	0.277	0.277	0.277	0.277
Segment WMT21	0.152	0.153	0.157	0.146	0.15	0.153
Segment WM22	0.398	0.399	0.4	0.398	0.4	0.402
Average	0.543	0.543	0.543	0.542	0.543	0.543

Table 3: Correlations over all metrics depending of the hyperparameters with a vocabulary size of 32,000

Vocabulary size	32000					
Tokenization algorithm	BPE			Unigram		
Beta	4	3	2	4	3	2
n-gram length	6					
System WMT20	0.875	0.876	0.877	0.876	0.876	0.877
System WMT21	0.715	0.716	0.717	0.716	0.717	0.718
System WMT22	0.837	0.835	0.83	0.837	0.835	0.83
Segment WMT20	0.277	0.277	0.277	0.277	0.277	0.277
Segment WMT21	0.157	0.158	0.157	0.158	0.158	0.153
Segment WM22	0.397	0.399	0.400	0.399	0.401	0.402
Average	0.543	**0.544**	0.543	**0.544**	**0.544**	0.543

Table 4: Tokengram_F results on the WMT20 dataset compared to chrF++

Language pair	tokengram_F system *r*	chrF++ system *r*	tokengram_F segment τ	chrF++ segment τ
en-cs	**0.865**	0.833	**0.485**	0.478
en-de	**0.961**	0.958	**0.371**	0.367
en-ru	**0.981**	0.952	**0.162**	0.156
en-ta	0.941	**0.956**	**0.590**	0.579
en-zh	0.851	**0.983**	**0.403**	0.388
en-ja	**0.949**	0.328	**0.521**	0.506
en-pl	**0.958**	0.315	**0.256**	0.255
en-iu	**0.433**	0.338	**0.340**	0.338
cs-en	**0.872**	0.844	**0.095**	0.09
de-en	0.997	**0.998**	**0.440**	0.435
pl-en	0.508	**0.970**	0.032	**0.034**
ta-en	**0.957**	0.522	0.184	**0.186**
km-en	**0.984**	0.965	**0.281**	0.275
ps-en	0.894	**0.964**	0.143	**0.145**
ja-en	**0.972**	0.763	**0.251**	0.245
ru-en	0.921	**0.977**	**0.055**	0.054
zh-en	**0.960**	0.841	**0.130**	**0.130**
iu-en	**0.765**	0.726	0.242	**0.246**

Table 5: Tokengram_F results on the WMT21 dataset compared to chrF (chrF was used in place of chrF++ as chrF++ results were not reported)

Language pair	tokengram_F system r	chrF system r	tokengram_F segment τ	chrF segment τ
en-cs	**0.978**	0.970	**0.549**	0.531
en-zh	**0.625**	0.549	**0.121**	0.092
en-ha	**0.760**	0.748	0.185	**0.186**
en-ja	**0.967**	0.966	**0.384**	0.371
en-ru	**0.756**	0.943	**0.214**	0.201
en-de	**0.842**	0.831	**0.448**	0.098
cs-en	**0.562**	**0.562**	**-0.052**	-0.053
zh-en	0.269	**0.723**	**0.395**	-0.035
ha-en	0.921	**0.924**	**0.021**	**0.021**
ja-en	0.823	**0.831**	**0.006**	0.005
ru-en	0.579	**0.593**	**-0.123**	-0.126
de-en	**0.424**	0.357	**-0.151**	-0.162
fr-de	**0.655**	0.646	0.049	**0.054**
de-fr	**0.504**	0.498	**0.111**	0.110
bn-hi	**0.949**	0.941	**0.079**	0.071
hi-bn	**0.877**	0.872	**0.335**	0.327
xh-zu	**0.999**	0.998	**0.306**	0.301
zu-xh	0.997	**0.999**	0.529	**0.530**

Table 6: Tokengram_F results on the WMT22 dataset compared to chrF++ (chrF was used in place of chrF++ as chrF++ results were not reported)

Language pair	tokengram_F system r	chrF system r	tokengram_F segment τ	chrF segment τ
en-cs	0.602	**0.689**	0.077	**0.147**
en-zh	**0.248**	0.210	-0.044	**0.051**
en-hr	0.899	**0.920**	**0.274**	0.185
en-ja	0.927	**0.931**	**0.241**	0.142
en-liv	**0.989**	0.988	**0.370**	0.101
en-ru	**0.852**	0.813	**0.659**	0.153
en-uk	0.869	**0.895**	**0.178**	0.177
en-de	0.799	**0.811**	**1.000**	0.085
liv-en	**0.985**	0.969	**0.500**	0.184
zh-en	0.787	**0.881**	**0.415**	0.071
sah-ruh	**1.000**	**1.000**	**0.856**	0.430
uk-cs	0.971	**0.979**	**0.350**	0.171
cs-uk	0.921	**0.927**	**0.311**	0.195

Embed_Llama: using LLM embeddings for the Metrics Shared Task

Sören Dréano
ML-Labs
Dublin City University
soren.dreano2@mail.dcu.ie

Derek Molloy
School of Electronic Engineering
Dublin City University
derek.molloy@dcu.ie

Noel Murphy
School of Electronic Engineering
Dublin City University
noel.murphy@dcu.ie

Abstract

Embed_llama is an assessment metric for language translation that hinges upon the utilization of the recently introduced Llama 2 Large Language Model (LLM), specifically focusing on its embedding layer, to transform sentences into a vector space that establishes connections between geometric and semantic proximities.

Investigations utilizing previous WMT datasets have revealed that within the Llama 2 architecture, relying solely on the initial embedding layer does not result in the highest degree of correlation when assessing machine translations. The incorporation of additional layers, however, holds the potential to augment the contextual understanding of sentences.

As a contribution to the WMT23 challenge, this study delves into the advantages derived from employing a pre-trained LLM that has not undergone fine-tuning specifically for translation evaluation tasks, to provide a metric conducive to operation on readily accessible consumer-grade hardware. This research digs into the observation that deeper layers within the model do not result in a linear increase in the spatial proximity between sentences within the vector space.

1 Introduction

The assessment of algorithm-generated translations entails the utilization of evaluation metrics that furnish quantitative scores to objectively gauge the precision of the model's output. Various methodologies are employed to juxtapose machine-generated translations with their human-generated counterparts, contingent upon the specific evaluation metric employed. In recent years, the realm of machine translation (MT) has witnessed notable advancements in terms of both translation accuracy and linguistic fluency.

Over time, there has been a significant enhancement in the correlation coefficient between human assessments and the automated evaluation of sentences generated by machines. Earlier metrics, such as BLEU (Papineni et al., 2002) or chrF++ (Popović, 2017), predominantly relied on the textual overlap between reference translations and the machine-generated counterparts. In contrast, contemporary approaches, exemplified by COMET (Rei et al., 2020), harness recent breakthroughs in Natural Language Processing(NLP) and transformer models, enabling them to consider not only individual words, but also to leverage contextual semantics for a more comprehensive evaluation.

In the domain of Machine Learning (ML) applied to NLP, the embedding layer assumes a pivotal role within neural network architectures, particularly in tasks centered on textual data. Its central objective lies in the transformation of discrete tokens, encompassing entities like words or characters into continuous vector representations. These vector representations, which maintain continuity, are amenable to acquisition and manipulation by neural networks and are commonly referred to as *word embeddings*.

2 Embed_Llama

The initial component in a Natural Language Processing (NLP) model is typically an embedding layer, which serves the purpose of converting the distinct identifiers of tokens within the input sentence into a vectorized representation. In this context, it is essential to emphasize that sentences conveying similar semantic content should exhibit proximity in the vector space, irrespective of the presence of word-level overlap, in contrast to sentences chosen randomly.

Embed_Llama draws inspiration from Word2vec (Mikolov et al., 2013) using Llama 2 (Touvron et al., 2023), a contemporary open-source pre-trained model. Rather than needing to train an extra NLP model for the purpose of assessing translation quality, a viable alternative approach involves uti-

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 738–745
December 6–7, 2023.

lizing a pre-trained, extensive neural network like Llama 2, which has been originally trained for next-token prediction. This approach allows for the investigation of how closely related sentences evolve across the model's various layers, all without incurring the supplementary expenses associated with fine-tuning or training anew.

2.1 Word2vec

Word2vec is a methodology that utilizes a neural network model to extract word associations from comprehensive textual corpora. Post training, this model possesses the capability to identify synonymous terms and offer word suggestions for unfinished sentences. As the terminology implies, Word2vec symbolizes individual words by employing distinct numerical vectors, systematically engineered to encapsulate both the semantic and syntactic characteristics inherent to the words.

Embed_Llama leverages vectorial space to estimate similarity or dissimilarity. This estimation is accomplished by computing the cosine distance between two sentences.

2.2 Vocabulary size

The lexical repertoire of Llama 2 encompasses 32,000 unique tokens, a figure lower than that of both GPT-2 (Radford et al., 2019) and GPT-NeoX (Black et al., 2022), which both employ 50,000 unique tokens. Regrettably, the specific vocabulary sizes pertaining to GPT-3 (Brown et al., 2020) and GPT-4 (OpenAI, 2023) remain undisclosed.

2.3 Embeddings

The dimensionality of the embedding space represents a hyperparameter subject to adjustment. Embeddings of higher dimensions have the capacity to capture more intricate relationships; however, it is noteworthy that such higher-dimensional embeddings may necessitate increased quantities of data and computational resources for their effective utilization.

2.4 Cosine distance

The cosine similarity metric quantifies the similarity between two n-dimensional vectors by computing the cosine of the angle between them. This scoring measure finds common application in the domain of text mining (Singhal, 2001). The general formula for two vectors A and B is:

$$\cos\theta = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

To enable efficient computation on consumer-grade hardware, the two sentences slated for comparison are padded to match the maximum token count of the longer sentence in the pair. Consequently, a sentence initially shaped as [length] transforms into the shape [length ×4,096] following its processing through the embedding layer.

3 Hyperparameter

As the objective of this current study revolves around the utilization of a pre-trained network, our sphere of influence is limited to a sole hyperparameter, namely, the number of blocks to retain before computing the cosine distance.

3.1 Block architecture

As shown in Figure 1, each block, denoted as the *LlamaDecoderLayer*, is structured with several components, including an attention layer, two normalization layers, and a multi-layer perceptron. The multi-layer perceptron, in turn, consists of three linear layers along with an associated activation function, while the attention layer also includes a rotary embedding layer.

The Llama 2 model, comprising 7 billion parameters, encompasses an embedding layer, 32 blocks, and a projection layer. To determine the optimal number of blocks, datasets extracted from the WMT challenge editions of 2020, 2021, and 2022 were employed. Due to the limited GPU memory allocation in the current project, it was only feasible to investigate the Llama 2 model up to a depth of 22 blocks, whereas the model has a total of 32 available blocks.

```
LlamaDecoderLayer(
  (self_attn): LlamaAttention(
    (q_proj): Linear(in_features=4096, out_features=4096, bias=False)
    (k_proj): Linear(in_features=4096, out_features=4096, bias=False)
    (v_proj): Linear(in_features=4096, out_features=4096, bias=False)
    (o_proj): Linear(in_features=4096, out_features=4096, bias=False)
    (rotary_emb): LlamaRotaryEmbedding()
  )
  (mlp): LlamaMLP(
    (gate_proj): Linear(in_features=4096, out_features=11008, bias=False)
    (up_proj): Linear(in_features=4096, out_features=11008, bias=False)
    (down_proj): Linear(in_features=11008, out_features=4096, bias=False)
    (act_fn): SiLUActivation()
  )
  (input_layernorm): LlamaRMSNorm()
  (post_attention_layernorm): LlamaRMSNorm()
)
```

Figure 1: Architecture of the Llama2 model as displayed in the Huggingface library

3.2 Exploring the depth

It was initially hypothesized that increasing the number of blocks would improve the contextual

representation of sentence meaning. Figure 2 reveals that, in contrast to our initial hypotheses, the augmentation of block quantities does not significantly modify the correlation between the systems and the ground truth, whether by augmentation or reduction. Furthermore, it is noteworthy that this correlation exhibits variability across different datasets. Specifically, the Pearson correlation between the number of layers and algorithm accuracy is 0.34 for the WMT20 dataset, but it decreases to -0.79 for the WMT22 dataset.

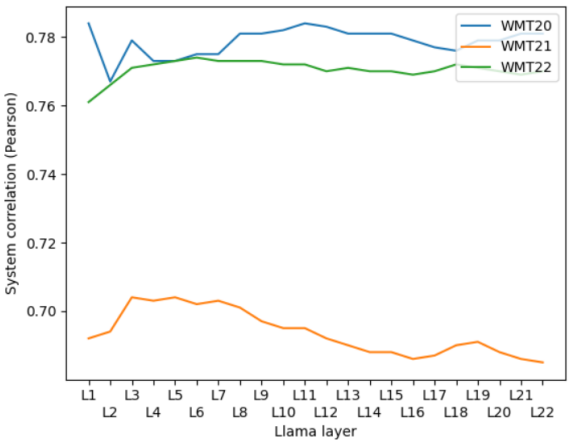


Figure 2: System correlations depending on the Llama layer

As shown in Figure 3, these observations hold with respect to segment correlation as well. The quantity of blocks employed in the Embed_Llama does not consistently enhance the metric's quality, whether for individual segments or entire systems.

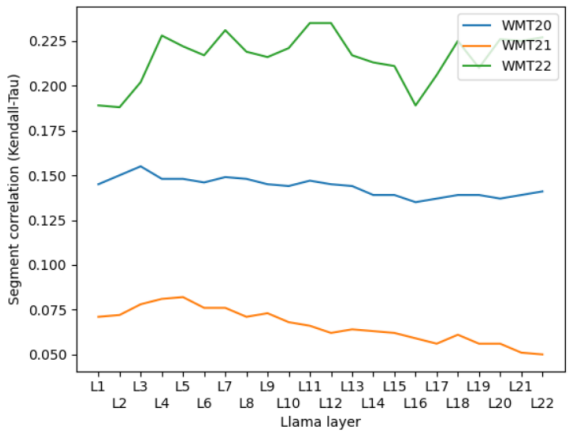


Figure 3: Segment correlations depending on the Llama layer

The highest levels of correlation between human rankings and Embed_Llama rankings were achieved by utilizing a mere two blocks following the embedding layer, resulting in optimal overall performance across the WMT20, WMT21, and WMT22 datasets. This not only expedited the computational process, but also decreased the GPU memory demands for metric computation.

3.3 Inter-languages variations

As depicted in Figure 4, the associations between metric quality and language pairs exhibit large variations. For instance, when considering the Hungarian-to-English language pair, the Pearson coefficient registers at 0.83, whereas it falls to -0.85 for the Czech-to-Ukrainian pair.

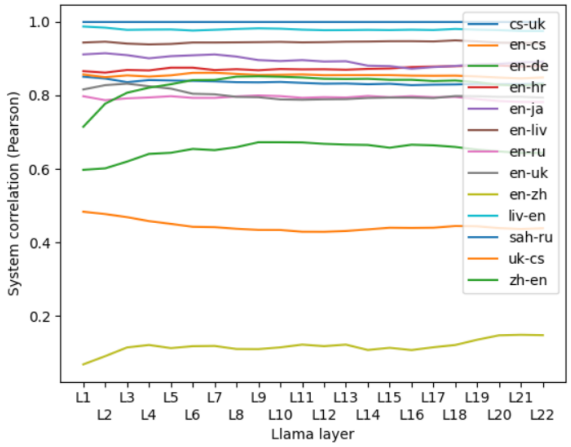


Figure 4: System correlations for each language pair in the WMT2022 dataset depending on the Llama layer

3.4 Intra-languages variations

Considerable variability is observed even within the same language pair across various datasets. For instance, the English-to-Chinese language pair is encompassed within the WMT2020, WMT2021, and WMT2022 datasets. However, as illustrated in Figure 5, no discernible correlations emerge between the number of utilized blocks and the quality of Embed_Llama scores. This is evident in the transformation of the Pearson coefficient, which shifts from -0.28 in the WMT2021 to 0.73 in the WMT2022 dataset.

3.5 Inter-datasets variations

Table 1 presents the average mean values and their corresponding standard deviations, showcasing the relationship between metric accuracy and the number of utilized blocks for language pairs common to all three datasets. It is notable that, apart for the English-to-Japanese pair, there exists a significant degree of variability in the performance of the

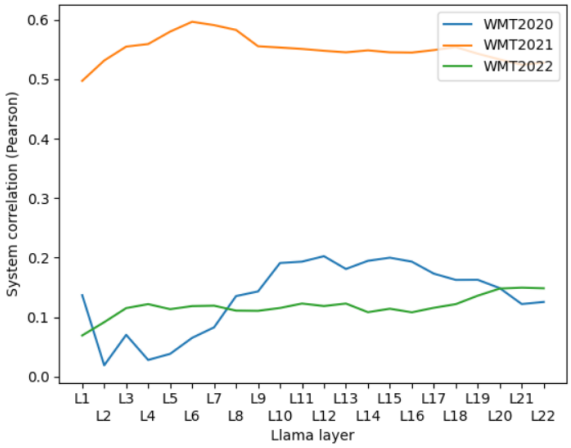


Figure 5: System correlations of the English-to-Chinese language pair depending on the Llama layer and the dataset

Language pair	Mean	Standard deviation
en-cs	-0.34	0.22
en-de	-0.02	0.4
en-ja	-0.89	0.04
en-ru	-0.02	0.54
en-zh	0.35	0.45
zh-en	-0.17	0.46

Table 1: Means and standard deviations of the system correlations depending on the Llama layer when WMT2020, WMT2021 and WMT2022 are merged

same language pairs across different datasets. This variability is underscored by the substantial standard deviations in relation to the absolute values of the means. Given the limited usage of just three datasets, it is essential to acknowledge that the relatively small sample size may hinder the ability to draw conclusive inferences regarding inter-dataset variability.

3.6 Full results

Tables 2, 4 and 6, correspondingly, present the Pearson correlation coefficients for datasets WMT2020, WMT2021 and WMT2021, illustrating the association between the algorithm-assigned scores and the actual rankings of the evaluated systems for individual language pairs.

With regard to segment-level correlations, they are presented in Tables 3, 5 and 7 for WMT2020, WMT2021, and WMT2022, respectively. It is noteworthy that, in contrast to system correlations, these are represented by Kendall coefficients, which are utilized as a measure of ordinal association.

It is noteworthy that the observed variations in these correlations are predominantly influenced by the specific language pairs, rather than the depth of the final block employed prior to cosine similarity computation, aligning with our anticipated outcome.

3.7 Source code

The source code of Embed_Llama is available at https://github.com/SorenDreano/embed_llama.

4 Conclusion

Although the authors initially anticipated that Embed_Llama would exhibit suboptimal performance for a majority of language pairs, except for those involving English, due to the apparent constraints posed by a limited vocabulary size and the nature of the dataset Llama 2 was trained on, the actual performance did not exhibit a significant underperformance.

The results from previous iterations of the WMT metrics shared task, presented in Appendix A, indicate that this approach may not meet the contemporary state-of-the-art standards exemplified by METRICX_XXL (unpublished) and COMET-22 (Rei et al., 2022).

The methodology involving the utilization of a non-finetuned, pre-trained Large Language Model (LLM) to assess translation quality through vector space similarity comparisons remains a prospective avenue of inquiry. This prospect gains relevance in light of forthcoming open-source models characterized by expansive vocabularies and training data encompassing diverse languages.

5 Further work

Given the recent proliferation of open-source LLMs, it is likely that another model, either presently or in the near future, may surpass the performance of Llama 2 for translation evaluation without necessitating any fine-tuning.

In the current investigation, the exploration has been confined to the 7 billion parameters model. It remains conceivable that employing a more extensive model with increased parameters may yield a more precise metric, albeit at the trade-off of heightened computational resource demands.

In the present evaluation, an exploration was limited to the initial 22 blocks. Subsequent endeavors

may consider augmenting this number, as doing so could potentially result in further benefits.

Moreover, it is worth noting that the optimal selection of the number of blocks to employ may be contingent upon the specific target language. Consequently, adjusting this hyperparameter based on the language in question could potentially yield enhanced correlation results.

In the scope of the current academic study, solely the cosine distance served as the chosen similarity measure for tensors. Future research endeavors may wish to investigate alternative distance metrics, such as the Euclidean distance or the Manhattan distance, for potential exploration.

6 Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. Gpt-neox-20b: An open-source autoregressive language model.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

OpenAI. 2023. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. Association for Computational Linguistics.

Amit Singhal. 2001. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24:35–43.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

A Appendix. Tables

Table 2: Pearson correlations (r) for the WMT20 system dataset depending on the Llama layer. The findings pertaining to the second layer are shown in bold, as it represents the prevailing default layer count within the Embed_Llama

Language pair	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11
en-cs	0.855	**0.866**	0.875	0.877	0.873	0.876	0.872	0.863	0.861	0.864	0.867
en-de	0.923	**0.928**	0.933	0.929	0.928	0.927	0.927	0.926	0.925	0.925	0.928
en-ru	0.856	**0.677**	0.769	0.772	0.81	0.829	0.821	0.902	0.912	0.899	0.918
en-ta	0.883	**0.901**	0.914	0.921	0.926	0.932	0.934	0.936	0.939	0.942	0.945
en-zh	0.137	**0.019**	0.07	0.028	0.038	0.065	0.083	0.135	0.143	0.191	0.193
en-ja	0.898	**0.892**	0.879	0.871	0.88	0.881	0.874	0.883	0.88	0.876	0.871
en-pl	0.88	**0.874**	0.88	0.878	0.875	0.864	0.868	0.873	0.873	0.872	0.868
en-iu	0.252	**0.226**	0.194	0.156	0.137	0.121	0.119	0.106	0.105	0.096	0.094
cs-en	0.795	**0.777**	0.752	0.771	0.789	0.788	0.796	0.795	0.796	0.789	0.775
de-en	0.992	**0.996**	0.996	0.994	0.99	0.99	0.99	0.988	0.988	0.988	0.99
pl-en	0.419	**0.403**	0.428	0.433	0.401	0.395	0.4	0.394	0.388	0.398	0.407
ta-en	0.876	**0.889**	0.918	0.922	0.919	0.923	0.921	0.919	0.917	0.919	0.918
km-en	0.954	**0.969**	0.988	0.988	0.988	0.989	0.99	0.989	0.987	0.986	0.985
ps-en	0.926	**0.874**	0.877	0.862	0.872	0.877	0.873	0.875	0.875	0.875	0.878
ja-en	0.913	**0.938**	0.947	0.946	0.946	0.948	0.949	0.948	0.946	0.948	0.949
ru-en	0.949	**0.939**	0.949	0.948	0.953	0.947	0.948	0.949	0.948	0.944	0.939
zh-en	0.97	**0.967**	0.963	0.956	0.957	0.955	0.955	0.954	0.954	0.951	0.951
iu-en	0.64	**0.676**	0.68	0.662	0.638	0.645	0.638	0.625	0.616	0.615	0.633

Language pair	L12	L13	L14	L15	L16	L17	L18	L19	L20	L21	L22
en-cs	0.865	0.866	0.86	0.858	0.851	0.856	0.855	0.857	0.866	0.874	0.874
en-de	0.925	0.926	0.922	0.921	0.917	0.918	0.918	0.921	0.923	0.929	0.929
en-ru	0.922	0.904	0.903	0.898	0.902	0.885	0.862	0.902	0.93	0.941	0.938
en-ta	0.942	0.942	0.944	0.946	0.947	0.947	0.948	0.948	0.948	0.946	0.946
en-zh	0.202	0.181	0.195	0.2	0.193	0.173	0.162	0.163	0.149	0.122	0.125
en-ja	0.87	0.867	0.869	0.865	0.865	0.857	0.858	0.848	0.838	0.836	0.836
en-pl	0.873	0.871	0.875	0.872	0.869	0.86	0.865	0.862	0.858	0.857	0.857
en-iu	0.084	0.085	0.082	0.084	0.083	0.078	0.073	0.082	0.091	0.092	0.086
cs-en	0.781	0.787	0.787	0.794	0.802	0.794	0.8	0.79	0.783	0.776	0.785
de-en	0.989	0.989	0.988	0.988	0.986	0.989	0.989	0.991	0.992	0.994	0.994
pl-en	0.408	0.42	0.421	0.418	0.417	0.414	0.413	0.41	0.406	0.415	0.415
ta-en	0.918	0.917	0.914	0.91	0.907	0.912	0.912	0.913	0.913	0.919	0.921
km-en	0.983	0.982	0.975	0.975	0.967	0.964	0.967	0.968	0.97	0.966	0.968
ps-en	0.876	0.881	0.887	0.892	0.891	0.888	0.889	0.889	0.889	0.895	0.897
ja-en	0.947	0.946	0.944	0.941	0.936	0.94	0.942	0.943	0.942	0.941	0.941
ru-en	0.938	0.938	0.938	0.939	0.94	0.935	0.935	0.935	0.933	0.935	0.937
zh-en	0.95	0.951	0.951	0.95	0.952	0.953	0.951	0.954	0.955	0.958	0.958
iu-en	0.619	0.613	0.606	0.6	0.602	0.627	0.63	0.639	0.644	0.662	0.66

Table 3: Kendall correlations (τ) for the WMT20 segment dataset depending on the Llama layer

Language pair	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11
en-cs	0.226	**0.231**	0.244	0.234	0.232	0.226	0.229	0.225	0.225	0.222	0.226
en-de	0.181	**0.19**	0.207	0.201	0.205	0.198	0.202	0.201	0.199	0.196	0.198
en-ru	0.028	**0.034**	0.037	0.045	0.038	0.045	0.042	0.044	0.043	0.046	0.05
en-ta	0.355	**0.354**	0.323	0.284	0.282	0.267	0.279	0.281	0.282	0.26	0.255
en-zh	0.147	**0.141**	0.164	0.144	0.138	0.141	0.147	0.152	0.152	0.16	0.165
en-ja	0.277	**0.281**	0.295	0.29	0.285	0.287	0.283	0.285	0.284	0.284	0.282
en-pl	0.097	**0.095**	0.103	0.101	0.101	0.102	0.098	0.101	0.098	0.094	0.102
en-iu	0.218	**0.224**	0.205	0.187	0.18	0.173	0.185	0.187	0.186	0.179	0.18
cs-en	0.064	**0.065**	0.073	0.072	0.077	0.076	0.076	0.073	0.074	0.078	0.083
de-en	0.349	**0.374**	0.387	0.384	0.39	0.389	0.385	0.376	0.372	0.375	0.381
pl-en	-0.016	**-0.016**	-0.0	-0.007	-0.0	0.0	-0.003	0.002	-0.002	-0.005	-0.006
ta-en	0.108	**0.118**	0.125	0.121	0.122	0.119	0.128	0.121	0.106	0.108	0.123
km-en	0.141	**0.145**	0.144	0.12	0.13	0.122	0.117	0.114	0.097	0.108	0.114
ps-en	0.088	**0.081**	0.074	0.076	0.058	0.053	0.06	0.061	0.065	0.061	0.08
ja-en	0.109	**0.136**	0.136	0.144	0.147	0.149	0.151	0.144	0.137	0.139	0.134
ru-en	0.011	**0.02**	0.019	0.022	0.026	0.022	0.031	0.023	0.02	0.02	0.017
zh-en	0.065	**0.067**	0.071	0.072	0.073	0.072	0.072	0.071	0.068	0.069	0.069
iu-en	0.16	**0.161**	0.175	0.178	0.187	0.191	0.195	0.194	0.195	0.198	0.199

Language pair	L12	L13	L14	L15	L16	L17	L18	L19	L20	L21	L22
en-cs	0.23	0.229	0.222	0.223	0.215	0.218	0.221	0.222	0.221	0.222	0.221
en-de	0.19	0.193	0.187	0.183	0.177	0.178	0.184	0.184	0.186	0.186	0.19
en-ru	0.043	0.037	0.043	0.042	0.045	0.036	0.039	0.037	0.039	0.043	0.044
en-ta	0.281	0.29	0.283	0.266	0.261	0.267	0.264	0.232	0.204	0.224	0.237
en-zh	0.166	0.168	0.162	0.165	0.16	0.163	0.166	0.163	0.165	0.166	0.164
en-ja	0.282	0.276	0.28	0.281	0.274	0.272	0.271	0.271	0.263	0.266	0.27
en-pl	0.098	0.096	0.091	0.087	0.093	0.093	0.095	0.093	0.09	0.092	0.092
en-iu	0.17	0.167	0.162	0.163	0.154	0.153	0.149	0.146	0.141	0.145	0.143
cs-en	0.081	0.073	0.073	0.076	0.074	0.079	0.081	0.081	0.077	0.073	0.075
de-en	0.377	0.377	0.371	0.372	0.361	0.37	0.37	0.378	0.374	0.383	0.386
pl-en	-0.004	0.0	-0.001	0.005	0.001	-0.001	0.001	-0.004	-0.006	-0.004	-0.003
ta-en	0.118	0.111	0.107	0.101	0.103	0.11	0.112	0.114	0.112	0.106	0.103
km-en	0.109	0.11	0.092	0.097	0.083	0.091	0.1	0.111	0.116	0.124	0.126
ps-en	0.074	0.081	0.067	0.068	0.064	0.056	0.066	0.07	0.07	0.067	0.07
ja-en	0.128	0.125	0.116	0.112	0.102	0.115	0.117	0.125	0.126	0.129	0.129
ru-en	0.015	0.007	0.009	0.008	0.013	0.012	0.015	0.018	0.02	0.025	0.022
zh-en	0.069	0.072	0.07	0.069	0.066	0.069	0.07	0.072	0.07	0.074	0.074
iu-en	0.192	0.177	0.173	0.176	0.173	0.179	0.184	0.183	0.188	0.189	0.189

Table 4: Pearson correlations (r) for the WMT21 system dataset depending on the Llama layer

Language pair	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11
en-cs	0.985	**0.984**	0.985	0.982	0.981	0.981	0.982	0.982	0.979	0.98	0.982
en-zh	0.497	**0.531**	0.555	0.559	0.58	0.597	0.591	0.583	0.555	0.553	0.551
en-ha	0.534	**0.551**	0.567	0.567	0.567	0.553	0.554	0.553	0.563	0.56	0.547
en-ja	0.82	**0.83**	0.838	0.836	0.83	0.84	0.837	0.839	0.829	0.822	0.816
en-ru	0.567	**0.621**	0.682	0.674	0.674	0.658	0.674	0.614	0.572	0.576	0.573
en-de	0.818	**0.793**	0.804	0.801	0.794	0.789	0.794	0.799	0.796	0.792	0.79
cs-en	0.542	**0.454**	0.435	0.428	0.429	0.432	0.426	0.426	0.429	0.42	0.423
zh-en	0.232	**0.186**	0.155	0.159	0.172	0.179	0.188	0.196	0.198	0.196	0.192
ha-en	0.825	**0.855**	0.858	0.855	0.867	0.868	0.867	0.867	0.865	0.86	0.855
ja-en	0.728	**0.726**	0.748	0.753	0.761	0.76	0.759	0.759	0.761	0.761	0.759
ru-en	0.613	**0.606**	0.535	0.514	0.5	0.496	0.502	0.506	0.506	0.504	0.509
de-en	0.171	**0.22**	0.237	0.238	0.221	0.23	0.223	0.223	0.225	0.234	0.243
fr-de	0.564	**0.556**	0.555	0.556	0.556	0.555	0.557	0.555	0.556	0.554	0.552
de-fr	0.477	**0.511**	0.578	0.579	0.579	0.58	0.579	0.575	0.563	0.564	0.574
bn-hi	0.908	**0.943**	0.94	0.942	0.937	0.929	0.941	0.95	0.943	0.941	0.942
hi-bn	0.879	**0.872**	0.913	0.93	0.925	0.915	0.912	0.907	0.911	0.908	0.915
xh-zu	0.952	**0.932**	0.934	0.931	0.923	0.909	0.904	0.905	0.898	0.899	0.891
zu-xh	0.95	**0.937**	0.97	0.973	0.975	0.972	0.971	0.972	0.976	0.976	0.976

Language pair	L12	L13	L14	L15	L16	L17	L18	L19	L20	L21	L22
en-cs	0.983	0.984	0.983	0.983	0.979	0.981	0.982	0.983	0.983	0.984	0.983
en-zh	0.548	0.545	0.548	0.545	0.545	0.549	0.554	0.543	0.533	0.524	0.528
en-ha	0.541	0.543	0.554	0.553	0.571	0.562	0.572	0.567	0.556	0.542	0.534
en-ja	0.81	0.804	0.795	0.793	0.781	0.783	0.786	0.777	0.772	0.774	0.779
en-ru	0.581	0.582	0.562	0.58	0.526	0.557	0.565	0.603	0.621	0.625	0.637
en-de	0.796	0.803	0.808	0.808	0.814	0.81	0.81	0.8	0.79	0.787	0.791
cs-en	0.394	0.386	0.383	0.387	0.388	0.395	0.41	0.425	0.422	0.433	0.435
zh-en	0.189	0.189	0.185	0.189	0.19	0.173	0.174	0.168	0.157	0.142	0.146
ha-en	0.846	0.842	0.84	0.841	0.841	0.827	0.828	0.829	0.823	0.818	0.814
ja-en	0.76	0.763	0.768	0.771	0.775	0.772	0.772	0.77	0.769	0.762	0.763
ru-en	0.51	0.512	0.513	0.52	0.521	0.52	0.529	0.534	0.535	0.532	0.528
de-en	0.23	0.217	0.212	0.196	0.184	0.207	0.21	0.205	0.203	0.215	0.214
fr-de	0.553	0.554	0.555	0.555	0.553	0.55	0.55	0.547	0.543	0.541	0.542
de-fr	0.566	0.564	0.558	0.558	0.556	0.567	0.564	0.561	0.563	0.574	0.577
bn-hi	0.935	0.932	0.931	0.931	0.93	0.933	0.933	0.938	0.94	0.93	0.922
hi-bn	0.904	0.892	0.876	0.86	0.869	0.863	0.859	0.865	0.864	0.852	0.829
xh-zu	0.9	0.903	0.906	0.903	0.909	0.905	0.908	0.912	0.903	0.898	0.896
zu-xh	0.977	0.979	0.981	0.981	0.981	0.981	0.977	0.977	0.976	0.969	0.962

Table 5: Kendall correlations (τ) for the WMT21 segment dataset depending on the Llama layer

Language pair	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11
en-cs	0.197	**0.198**	0.243	0.243	0.238	0.223	0.221	0.219	0.219	0.214	0.22
en-zh	0.028	**0.028**	0.048	0.042	0.042	0.044	0.039	0.041	0.039	0.037	0.037
en-ha	0.073	**0.077**	0.087	0.081	0.077	0.071	0.071	0.068	0.062	0.061	0.062
en-ja	0.181	**0.184**	0.204	0.21	0.214	0.204	0.197	0.196	0.204	0.196	0.192
en-ru	0.094	**0.101**	0.099	0.099	0.089	0.09	0.091	0.086	0.081	0.087	0.084
en-de	0.241	**0.31**	0.172	0.172	0.241	0.31	0.31	0.241	0.241	0.241	0.241
cs-en	-0.042	**-0.048**	-0.049	-0.045	-0.041	-0.044	-0.042	-0.05	-0.046	-0.045	-0.051
zh-en	0.319	**0.286**	0.319	0.384	0.395	0.319	0.308	0.319	0.384	0.297	0.276
ha-en	-0.022	**-0.016**	-0.007	-0.003	-0.003	-0.009	-0.008	-0.007	-0.011	-0.014	-0.011
ja-en	-0.028	**-0.021**	-0.019	-0.018	-0.015	-0.017	-0.014	-0.015	-0.014	-0.012	-0.015
ru-en	-0.121	**-0.124**	-0.12	-0.12	-0.117	-0.123	-0.12	-0.119	-0.121	-0.124	-0.122
de-en	-0.155	**-0.158**	-0.151	-0.148	-0.153	-0.157	-0.155	-0.156	-0.158	-0.155	-0.152
fr-de	0.057	**0.058**	0.055	0.052	0.044	0.055	0.062	0.057	0.053	0.047	0.051
de-fr	0.054	**0.066**	0.082	0.079	0.075	0.051	0.057	0.054	0.049	0.055	0.055
bn-hi	-0.006	**0.007**	0.016	0.026	0.026	0.024	0.024	0.022	0.015	0.023	0.014
hi-bn	0.132	**0.119**	0.126	0.135	0.139	0.137	0.145	0.15	0.147	0.146	0.14
xh-zu	0.128	**0.129**	0.139	0.13	0.118	0.104	0.101	0.101	0.09	0.082	0.065
zu-xh	0.169	**0.139**	0.181	0.181	0.16	0.129	0.122	0.116	0.12	0.118	0.109

Language pair	L12	L13	L14	L15	L16	L17	L18	L19	L20	L21	L22
en-cs	0.213	0.218	0.213	0.22	0.198	0.209	0.209	0.196	0.193	0.194	0.197
en-zh	0.042	0.045	0.048	0.048	0.043	0.043	0.043	0.046	0.041	0.04	0.04
en-ha	0.058	0.06	0.061	0.061	0.061	0.06	0.062	0.064	0.068	0.067	0.067
en-ja	0.189	0.187	0.196	0.193	0.202	0.196	0.195	0.19	0.184	0.179	0.179
en-ru	0.088	0.089	0.109	0.107	0.116	0.116	0.109	0.106	0.108	0.104	0.105
en-de	0.172	0.172	0.172	0.172	0.034	0.034	0.034	0.034	0.034	0.034	0.034
cs-en	-0.047	-0.048	-0.045	-0.055	-0.054	-0.06	-0.056	-0.055	-0.053	-0.058	-0.058
zh-en	0.297	0.319	0.265	0.265	0.33	0.286	0.362	0.286	0.297	0.276	0.276
ha-en	-0.014	-0.013	-0.011	-0.014	-0.012	-0.015	-0.014	-0.014	-0.014	-0.014	-0.014
ja-en	-0.016	-0.02	-0.021	-0.02	-0.02	-0.022	-0.022	-0.021	-0.02	-0.021	-0.018
ru-en	-0.121	-0.121	-0.123	-0.123	-0.123	-0.126	-0.124	-0.116	-0.116	-0.12	-0.121
de-en	-0.152	-0.152	-0.152	-0.154	-0.159	-0.154	-0.154	-0.155	-0.156	-0.155	-0.157
fr-de	0.049	0.047	0.043	0.041	0.031	0.034	0.032	0.036	0.032	0.032	0.031
de-fr	0.039	0.04	0.035	0.039	0.049	0.046	0.053	0.046	0.046	0.046	0.052
bn-hi	0.015	0.02	0.022	0.022	0.019	0.02	0.023	0.023	0.027	0.025	0.025
hi-bn	0.123	0.114	0.114	0.113	0.114	0.12	0.12	0.116	0.108	0.091	0.074
xh-zu	0.076	0.083	0.081	0.094	0.096	0.088	0.098	0.089	0.085	0.083	0.07
zu-xh	0.118	0.122	0.14	0.131	0.141	0.138	0.141	0.144	0.137	0.125	0.122

Table 6: Pearson correlations (r) for the WMT22 system dataset depending on the Llama layer

Language pair	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11
en-cs	0.484	**0.477**	0.469	0.458	0.451	0.443	0.442	0.437	0.434	0.434	0.429
en-zh	0.069	**0.091**	0.115	0.122	0.113	0.118	0.119	0.111	0.11	0.115	0.123
en-hr	0.866	**0.862**	0.869	0.868	0.875	0.875	0.869	0.871	0.868	0.871	0.87
en-ja	0.911	**0.914**	0.909	0.901	0.906	0.909	0.911	0.905	0.896	0.893	0.896
en-liv	0.943	**0.946**	0.941	0.938	0.94	0.943	0.943	0.944	0.945	0.945	0.944
en-ru	0.798	**0.787**	0.792	0.794	0.798	0.793	0.793	0.797	0.8	0.798	0.793
en-uk	0.816	**0.828**	0.832	0.825	0.818	0.804	0.803	0.796	0.795	0.789	0.788
en-de	0.598	**0.602**	0.62	0.641	0.644	0.654	0.651	0.659	0.673	0.672	0.672
liv-en	0.987	**0.984**	0.978	0.979	0.979	0.976	0.978	0.98	0.982	0.981	0.978
zh-en	0.715	**0.777**	0.807	0.821	0.83	0.841	0.842	0.851	0.852	0.85	0.848
sah-ru	1.0	**1.0**	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
uk-cs	0.857	**0.849**	0.854	0.851	0.854	0.861	0.861	0.858	0.855	0.856	0.857
cs-uk	0.851	**0.846**	0.836	0.841	0.84	0.839	0.838	0.836	0.835	0.836	0.834

Language pair	L12	L13	L14	L15	L16	L17	L18	L19	L20	L21	L22
en-cs	0.429	0.431	0.436	0.44	0.44	0.44	0.445	0.444	0.439	0.437	0.439
en-zh	0.119	0.123	0.108	0.114	0.108	0.116	0.122	0.136	0.148	0.149	0.148
en-hr	0.871	0.87	0.872	0.873	0.877	0.878	0.879	0.881	0.881	0.879	0.879
en-ja	0.892	0.893	0.881	0.879	0.872	0.877	0.88	0.884	0.885	0.889	0.891
en-liv	0.944	0.946	0.946	0.947	0.947	0.946	0.949	0.947	0.943	0.942	0.942
en-ru	0.795	0.794	0.798	0.795	0.798	0.795	0.796	0.789	0.785	0.782	0.782
en-uk	0.789	0.79	0.793	0.794	0.794	0.793	0.798	0.798	0.796	0.796	0.794
en-de	0.668	0.666	0.665	0.658	0.666	0.664	0.66	0.652	0.648	0.644	0.644
liv-en	0.977	0.977	0.978	0.977	0.978	0.977	0.98	0.978	0.977	0.975	0.975
zh-en	0.845	0.844	0.845	0.842	0.842	0.839	0.84	0.836	0.828	0.826	0.83
sah-ru	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
uk-cs	0.855	0.855	0.855	0.855	0.854	0.853	0.853	0.851	0.848	0.846	0.848
cs-uk	0.832	0.832	0.83	0.831	0.828	0.829	0.83	0.831	0.83	0.83	0.832

Table 7: Kendall correlations (τ) for the WMT22 segment dataset depending on the Llama layer

Language pair	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11
en-cs	0.014	**0.014**	0.023	0.024	0.021	0.024	0.021	0.021	0.015	0.009	0.013
en-zh	-0.041	**-0.03**	-0.034	-0.039	-0.035	-0.024	-0.026	-0.027	-0.026	-0.029	-0.026
en-hr	0.119	**0.115**	0.14	0.125	0.123	0.11	0.117	0.118	0.128	0.123	0.122
en-ja	0.097	**0.098**	0.101	0.091	0.085	0.082	0.081	0.082	0.079	0.08	0.075
en-liv	0.362	**0.332**	0.35	0.34	0.33	0.33	0.334	0.31	0.292	0.322	0.332
en-ru	0.262	**0.227**	0.234	0.248	0.23	0.199	0.188	0.227	0.223	0.244	0.234
en-uk	0.081	**0.063**	0.062	0.062	0.066	0.055	0.044	0.041	0.04	0.048	0.052
en-de	0.4	**0.4**	0.4	0.8	0.8	0.8	1.0	0.8	0.8	0.8	1.0
liv-en	0.247	**0.276**	0.311	0.303	0.302	0.302	0.286	0.298	0.291	0.306	0.314
zh-en	0.179	**0.217**	0.241	0.217	0.195	0.225	0.223	0.213	0.193	0.209	0.203
sah-ru	0.458	**0.419**	0.484	0.492	0.478	0.45	0.456	0.47	0.461	0.456	0.444
uk-cs	0.145	**0.159**	0.16	0.153	0.141	0.125	0.132	0.146	0.167	0.156	0.146
cs-uk	0.133	**0.149**	0.159	0.154	0.154	0.148	0.15	0.151	0.149	0.154	0.151

Language pair	L12	L13	L14	L15	L16	L17	L18	L19	L20	L21	L22
en-cs	0.008	0.014	0.014	0.011	0.02	0.018	0.017	0.011	0.011	0.008	0.008
en-zh	-0.03	-0.028	-0.034	-0.033	-0.031	-0.033	-0.033	-0.028	-0.029	-0.027	-0.027
en-hr	0.118	0.11	0.119	0.116	0.109	0.101	0.106	0.118	0.127	0.129	0.128
en-ja	0.081	0.083	0.082	0.083	0.081	0.083	0.084	0.075	0.072	0.074	0.074
en-liv	0.344	0.344	0.346	0.34	0.328	0.328	0.346	0.34	0.34	0.344	0.348
en-ru	0.22	0.202	0.174	0.188	0.16	0.167	0.167	0.202	0.192	0.195	0.206
en-uk	0.049	0.047	0.048	0.048	0.044	0.049	0.042	0.036	0.049	0.038	0.034
en-de	1.0	0.8	0.8	0.8	0.6	0.8	1.0	0.8	1.0	1.0	1.0
liv-en	0.313	0.286	0.261	0.26	0.249	0.258	0.269	0.28	0.272	0.277	0.286
zh-en	0.225	0.219	0.195	0.191	0.179	0.185	0.179	0.183	0.211	0.213	0.209
sah-ru	0.441	0.436	0.45	0.439	0.43	0.427	0.447	0.416	0.408	0.399	0.408
uk-cs	0.144	0.15	0.155	0.153	0.149	0.159	0.158	0.156	0.143	0.137	0.137
cs-uk	0.148	0.152	0.154	0.15	0.14	0.14	0.142	0.141	0.139	0.141	0.141

eBLEU: Unexpectedly Good Machine Translation Evaluation Using Simple Word Embeddings

Muhammad ElNokrashy
Microsoft
Cairo, Egypt
muelnokr@microsoft.com

Tom Kocmi
Microsoft
Munich, Germany
tomkocmi@microsoft.com

Abstract

We propose eBLEU, a metric inspired by BLEU metric that uses embedding similarities instead of string matches. We introduce *meaning diffusion vectors* to enable matching n-grams of semantically similar words in a BLEU-like algorithm, using efficient, non-contextual word embeddings like fastText. On WMT23 data, eBLEU beats BLEU and ChrF by around 3.8% system-level score, approaching BERTScore at -0.9% absolute difference. In WMT22 scenarios, eBLEU outperforms f101spBLEU and ChrF in MQM by $2.2\%-3.6\%$. Curiously, on MTurk evaluations, eBLEU surpasses past methods by $3.9\%-8.2\%$ (f200spBLEU, COMET-22). eBLEU presents an interesting middle-ground between traditional metrics and pretrained metrics.

1 Introduction

The machine translation field has improved significantly, with various metrics developed to measure translation quality. Translation quality in human eyes is usually a delicate balance to convey meaning, style, tone, and other dimensions of text from one language into another with different idioms and concept ontologies. After all, translation is not only about translating words from one language to another literally, but ensuring that the core meaning behind is also accurately conveyed.

Traditional metrics, like the **BLEU** score (Papineni et al., 2002) or ChrF (Popović, 2015), have proven effective over last 20 year. However, there has been growing evidence that they have not kept pace with the performance of recent NMT and LLM MT systems (Kocmi et al., 2021; Freitag et al., 2022). BLEU essentially computes a score based on string n-grams matches. One clear limitation of this approach is that it fails to recognize semantically similar words. For instance, in the eyes of BLEU, the words (*cat*, *kitty*) are as different as (*cat*, *book*) or (*fire*, *water*).

Recent neural metrics, on the other hand, have explored the potential of leveraging pretrained language models for encoding entire sentences. These models either compare encoded sentences in a shared embedding space or employ a trained classifier to predict human scores, as demonstrated by Rei et al. (2020); Zhang et al. (2020); Freitag et al. (2022). These methods are more capable of capturing semantic nuances.

In this paper, we introduce **eBLEU**, a metric designed to address the mentioned limitation of the BLEU score while keeping the calculation as close to BLEU as possible by using the word embedding similarities instead of string matching. By doing so, eBLEU enhances the metric by recognizing semantically similar n-grams. Our method relies on the *meaning diffusion map* to approximate n-gram matching in a BLEU-like algorithm. The core implementation leverages efficient, non-contextual word embeddings, such as fastText embeddings.

2 Related work

In machine translation, measuring quality is a balance of many potentially competing factors. The most prominent are language quality (fluency) and accuracy of meaning conveyed (adequacy). Other factors may be critical in special scenarios. Consider the conveyance of tone or cultural register in translated dialog (see for example registers in East-Asian languages). Or the conveyance of flow in a translated play (see some examples of translations of the Greek epic Iliad in Mendelsohn, 2011).

Traditional automatic quality assessment methods, like BLEU and METEOR (Banerjee and Lavie, 2005), rely on string matching against a reference. The more matches, the more a candidate captures of the intended meaning in the reference, as proxy for adequacy. While features like n-gram matching in BLEU and explicit ordering penalties in METEOR act as proxy for fluency.

Such metrics suffer from limitations inherent to

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 746–750
December 6–7, 2023.

their literal string matching core, which some try to mitigate (e.g. via lemmatization or synonym dictionaries). These limitations are clearer in light of the more complex and semantically rich language produced by recent Neural MT systems and Large Language Model MT systems.

BERTScore (Zhang et al., 2020) utilizes a similar idea to ours, matching contextual encodings of words in candidate/reference pairs. While it uses unigrams only, eBLEU uses n-grams as well, and calculates token matches differently. Other systems, like COMET (Rei et al., 2020), are finetuned on human judgement scores for machine translation evaluation specifically.

3 Preliminaries

3.1 BLEU

The BLEU formula applied to a single candidate/reference sentence pair X, Y is:

$$\underset{N}{\texttt{BLEU}}(X,Y) = \texttt{bp}(X,Y) \prod_{n \in 1..N} \texttt{pr}_n(X,Y)^{w_n} \tag{1}$$

where $\texttt{bp}(X,Y)$ is the brevity penalty. This score ranges between $[0,1]$ for lowest and highest match.

The n-gram precision $\texttt{pr}_n(X,Y)$ is:

$$\frac{\sum_{s \in [X]^n} \min(C(s,X), C(s,Y))}{\sum_{s \in [X]^n} C(s,X)} \tag{2}$$

Set of n-gram substrings in candidate

Count of s in candidate

3.2 Embeddings

At the core, our method utilizes simple word embeddings that can be generated from sub-word information or memorized for full words as appropriate. We do not require tokenization of words into sub-words. Here we use the fastText word embeddings (Bojanowski et al., 2017). Other simple word embeddings should be appropriate as-is but were not tested. FastText is trained for every language separately and we require a trained fastText model for the target language in any translation pair.

3.3 String Matching

Strings under strict equality are literal representations of unique identities: the string *abc* is equal only to *abc* itself. This works for BLEU. Now we want to match based on the closeness of meaning instead, where (*cat*, *cats*) would be closer together than (*cat*, *book*).

4 eBLEU description

We propose the following formulation for an embedding-based matching in the style of BLEU precision from eq. (2).

Let X refer to the candidate sentence, and Y refer to the reference sentence. Now, given an asymmetric similarity function $\texttt{mdSim}(a \mid b)$ from a with reference to b, we can define the following analogous values for "precision" and "recall":

$$\text{precision: } \texttt{pr}(X,Y) = \texttt{mdSim}(Y \mid X) \tag{3}$$

$$\text{recall: } \texttt{re}(X,Y) = \texttt{mdSim}(X \mid Y) \tag{4}$$

$\overline{m}_n$ is the n-gram score of the pair, defined as the geometric mean of the n-gram precision and recall from eq. (3, 4):

$$\overline{m}_n = \left(\texttt{pr}_n(X,Y) \cdot \texttt{re}_n(X,Y)\right)^{\frac{1}{2}} \tag{5}$$

The final score is a weighted geometric average of the n-gram-based scores $\overline{m}_n$ between candidate and reference, for $N = 4$ and $w_n = N - n$.

$$\underset{N}{\texttt{eBLEU}}(X,Y) = \texttt{lp}(X,Y) \prod_{n \in 1..N} \overline{m}_n^{\,w_n/N} \tag{6}$$

where $\texttt{lp}$ is a modified length penalty which penalizes longer candidates as well.

(e)BLEU This shows the analogous structure of eBLEU compared to BLEU, given an appropriate definition for $\texttt{mdSim}$ as used in eq. (3, 4).

4.1 Aggregating Similarity Values

Similar to $\texttt{pr}_n(X,Y)$, we want $\texttt{mdSim}(Y \mid X)$ to be a single value for a candidate/reference sentence pair, as if aggregating the meaning diffusion values m_x for $x \in X$:

$$\texttt{mdSim}(Y \mid X) = \frac{\sum_x \min(\mathbf{m}_X, \mathbf{m}_{X|Y})}{\sum_x \mathbf{m}_X} \tag{7}$$

Index over words in cand.

MD vector of words in cand.

Compare Equation (7) for eBLEU with Equation (2) for BLEU.

4.2 Meaning Diffusion

Meaning Diffusion (MD) is a value for each word in a sentence indicating *the ratio of similar words*

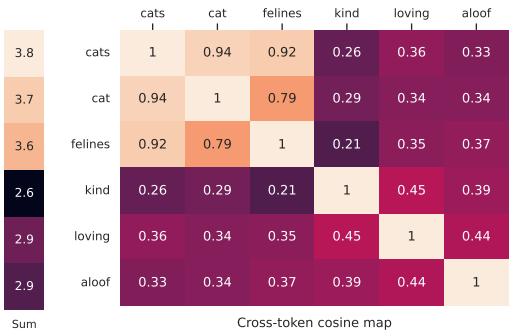


Figure 1: *(Right)* Meaning Diffusion Map between some words for illustration. Notice the similarity of *cats, cat, felines* and the relative similarity of *kind, loving* but not *aloof*. *(Left)* Meaning Diffusion Vector is the sum of a word's similarity to all words in the sentence.

in the same sentence. This allows claims such as "*there exists* $2/7$ `eats` " in:

The cat `eats` *, no,* `devours` *the food.*

That is: $m_{\text{eats}} \approx 2/7$. See also Figure 1 *(Left)*.

MD Map $\hat{\mathbf{S}}_{y,y'}$ is a weighted sum over the candidate side (x) with `softmax`-normalized weights. It approximates the similarity matrix of the reference Y against itself (see Figure 1 *(Right)*), as seen *through* the candidate X. The **L1** variant replaces `softmax` with a simple division by the sum of values: $\mathbf{S}_{y,x} / \sum_x \mathbf{S}_{y,x}$.

MD Vectors $\mathbf{m}_*$ represent each word's total closeness to all other words in the same sentence ($\mathbf{m}_Y$) or through a candidate sentence ($\mathbf{m}_{Y|X}$).

$$\hat{\mathbf{S}}_{y,y'} = \underset{x}{\texttt{softmax}}(\mathbf{S}_{y,x}) \cdot \mathbf{S}_{x,y} \tag{8}$$

$$\mathbf{m}_{Y|X} = \sum_{y'} \hat{\mathbf{S}}_{y,y'} \tag{9}$$

$$\mathbf{m}_Y = \sum_{y} \mathbf{S}_{y,y} \tag{10}$$

Vector Similarity For the candidate/reference X, Y, let $\mathbf{X}, \mathbf{Y}$ be the embedding matrices shaped as *token* $\times$ *embedding*. $\mathbf{S}_{x,y}$ is Cosine vector similarity clipped within $[0, 1]$, defined as:

$$\mathbf{S}_{x,y} = \texttt{clip}_{[0,1]} \cos_{\textit{embedding}}(\mathbf{X}, \mathbf{Y}^\top) \tag{11}$$

4.3 N-gram Scores

For each $n \in 1..N$, we calculate the n-gram score of a sentence pair using $\mathbf{S}^n_{x,y}$: the geometric mean of the cosine scores of adjacent words in each sentence, such that the n-gram-aware $\mathbf{S}^n_{x,y}$ is of shape $|X| - n + 1 \times |Y| - n + 1$.

4.4 Length Penalty

The length penalty penalizes length mismatch between candidate and reference, as used in eq. (6):

$$\texttt{lp}(X, Y) = \begin{cases} 1.0 & \texttt{ratio} \leq 0.5 \\ e^{0.5 - \texttt{ratio}} & \text{else} \end{cases} \tag{12}$$

$$\texttt{ratio} = \frac{\texttt{abs}(|X| - |Y|)}{|Y|} \tag{13}$$

5 Evaluation and Results

In this section, we describe the evaluation of the metric and the results

5.1 Meta-evaluation

We use the WMT Metrics 2022 test set (Freitag et al., 2021) which contains human judgments based on three different protocols: MQM, DA+SQM and MTurk DA. The translation systems are mainly from participants of the WMT22 General MT shared task (Kocmi et al., 2022). The source segments and human reference translations for each language pair contain around 2,000 sentences from four different texts domains: news, social, conversational, and e-commerce.

Human labels are produced via three methods:

- MQM - annotated by professionals who mark individual errors in each translation, as described in (Freitag et al., 2021)
- DA+SQM - professional annotators are asked to rate each translation on a scale 0-100 (Kocmi et al., 2022)
- MTurk DA - low paid crowd of MTurk annotators is asked to rate each translation on a scale 0-100, for how much it resembles human reference (Kocmi et al., 2022)

To determine the correlation of automatic metrics with humans, we measure system-level, pairwise accuracy (Kocmi et al., 2021), which is defined as the number of system pairs ranked correctly by the metric with respect to the human ranking divided by the total number of system pair comparisons. Formally:

$$\text{Accuracy} = \frac{|\text{sign}(\text{metric}\Delta) == \text{sign}(\text{human}\Delta)|}{|\text{all system pairs}|}$$

We reproduced scores reported in the WMT22 Metrics shared task findings paper with the official

System	MQM	DA-SQM	MTurk
COMET-22	**83.94%**	**84.19%**	62.61%
COMET-20	83.58%	82.17%	63.53%
BERTScore	77.37%	75.92%	66.57%
f101spBLEU	**74.45%**	74.26%	65.96%
f200spBLEU	74.09%	74.26%	**66.87%**
chrF	73.36%	**75.92%**	66.57%
BLEU	70.80%	70.22%	65.35%
eBLEU-FastText			
↪ L1	**76.64%**	74.45%	68.69%
↪ Softmax*	74.82%	72.43%	**70.82%**

Table 1: System-level WMT22 results on 3 human labeling scenarios. The Softmax variant of eBLEU was submitted to WMT23.

System	Rank	Score
COMET	2	**93.5%**
BERTScore	7	90.2%
f200spBLEU	11	**86.8%**
BLEU	12	85.9%
ChrF	12	85.2%
eBLEU-FastText		
↪ Softmax	9	**89.3%**

Table 2: WMT23 System-level ranking clusters and correlations on en-de, he-en, zh-en language pairs.[2]

WMT22 script.[1] Scores match Table 8 (DA+SQM and DA) and Table 11 for MQM of the WMT22 Metrics findings paper (Freitag et al., 2022).

5.2 Results

On **WMT23** scenarios (Table 2), eBLEU scores $\mathbf{89.3}\%$, improving noticeably on ChrF, BLEU, and f200spBLEU, beating the latter by $\mathbf{2.5}\%$ points. Its ranking cluster (**9**) puts it much closer to more sophisticated embedding-based metrics (like BERTScore) than string metrics like BLEU. Notably, this was achieved by the Softmax variant, which scored below the L1 variant on the more accurate human MQM and DA-SQM scenarios.

On **WMT22** scenarios (Table 1), eBLEU outperforms both f101spBLEU and ChrF in MQM by $2.2\% - 3.6\%$ in system-level accuracy.

eBLEU shows SOTA correlation with MTurk evaluations at $\mathbf{70.82}\%$, beating existing methods by $3.9\% - 8.2\%$ (f200spBLEU, COMET-22). Although Freitag et al. (2022) shows them to be of sub-optimal quality, this is interesting as MTurk evaluations often involve manual n-gram matching—a nice result given the intuition behind our method.

6 Conclusion

6.1 eBLEU: Between Strings and Neural Eval

In this paper, we introduced eBLEU, a novel metric that adapts the BLEU algorithm by adding embedding-based semantic understanding. By incorporating word embedding similarities and leveraging *meaning diffusion vectors*, eBLEU bridges the gap between literal and semantic matching.

We show that eBLEU can outperform widely adopted metrics like (sp)BLEU and ChrF, and approaches some pretrained contextual embedding-based metrics, like BERTScore, using simpler, cheaper-to-compute embeddings like fastText.

On WMT23, eBLEU scores 89.3%, placing almost halfway between BLEU, and COMET, an especially finetuned model for MT evaluation.

Although eBLEU lags behind the latest pretrained metrics, it presents an interesting approach for a simple semantically informed metric.

6.2 Limitations

However, it is important to recognize the limitations. Fundamentally, eBLEU does not attempt to improve the BLEU formula as a proxy for adequacy and fluency. Thus predictably, it lags far behind the latest pretrained metrics such as COMET or BLEURT. As language models, the core of these systems holds the advantages of large pre-training data, contextual understanding of input candidates and references, and potentially task-specific fine-tuning for the translation domain. Their more general nature allows for much improved measurement of adequacy and fluency among the range of possible translations that humans may produce and judge acceptable.

In summary, eBLEU offers a semantically-aware machine translation evaluation metric extending standardized BLEU algorithm. There may exist other such methods that bridge the gap further while improving inference time, efficiency, or interpretability where needed.

[1] https://github.com/google-research/mt-metrics-eval
[2] As provided by the WMT team.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics. [Cited on page 1.]

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146. [Cited on page 2.]

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474. [Cited on page 3.]

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics. [Cited on pages 1 and 4.]

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics. [Cited on page 3.]

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics. [Cited on pages 1 and 3.]

Daniel Mendelsohn. 2011. Englishing the iliad: Grading four rival translations. [Cited on page 1.]

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics. [Cited on page 1.]

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics. [Cited on page 1.]

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics. [Cited on pages 1 and 2.]

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. [Cited on pages 1 and 2.]

COMETOID: Distilling Strong Reference-based Machine Translation Metrics Into Even Stronger Quality Estimation Metrics

Thamme Gowda and **Tom Kocmi** and **Marcin Junczys-Dowmunt**
Microsoft Translator
Redmond, WA, USA
{thammegowda,tomkocmi,marcinjd}@microsoft.com

Abstract

This paper describes our submissions to the 2023 Conference on Machine Translation (WMT23) Metrics shared task. Knowledge distillation is commonly used to create smaller student models that mimic a larger teacher model while reducing the model size and hence inference cost in production. In this work, we apply knowledge distillation to machine translation evaluation metrics and distill existing reference-based teacher metrics into reference-free (quality estimation; QE) student metrics. We mainly focus on students of Unbabel's `COMET22` reference-based metric. When evaluating on the official WMT22 Metrics evaluation task, our distilled `Cometoid` QE metrics outperform all other QE metrics on that set while matching or out-performing the reference-based teacher metric. Our metrics never see the human ground-truth scores directly – only the teacher metric was trained on human scores by its original creators. We also distill `ChrF` sentence-level scores into a neural QE metric and find that our reference-free (and fully human-score-free) student metric `ChrFoid` outperforms its teacher metric by over 7% pairwise accuracy on the same WMT22 task, rivaling other existing QE metrics.[1]

1 Introduction

The Conference on Machine Translation (WMT) organizes an annual shared task for meta-evaluation of machine translation (MT) evaluation metrics (Freitag et al., 2022), where numerous MT evaluation metrics are proposed and revised each year. The MT metrics are broadly categorized as: (i) reference-based metrics, which score MT hypothesis against one or more reference translations from humans, and (ii) reference-free metrics, which do not require references and instead score hypothesis directly against the source sentence. Reference-free metrics, also known as quality estimation (QE) metrics, are an attractive choice in scenarios where reference translations are either unavailable or unreliable. However, currently, QE metrics lag behind the reference-based metrics by a considerable margin according to metrics meta-evaluation results (Freitag et al., 2022).

Knowledge distillation (KD) (Liang et al., 2008; Hinton et al., 2015) is commonly used to create smaller student models that mimic larger teacher models (Kim and Rush, 2016) which reduces computational cost when deploying models in production (Kim et al., 2019). Other use cases of KD in MT include distillation from auto-regressive teacher translation models to non-autoregressive students (Zhou et al., 2020) where the students "suffer" from an information bottleneck (here: no access to their own previous output in a time sequence) which impedes their performance when trained on original data. The simplified and probably smoothed output distribution of the teacher is easier to "digest" and often results in improved performance for the student.

In this work, we treat existing reference-based metrics as teachers and by applying knowledge distillation, we create reference-free student metrics that completely eliminate the need for references in evaluation. This is achieved by introducing a hard information bottleneck: just dropping the reference during training while keeping the original reference-based teacher score.

2 Experiments

2.1 Data Preparation

Our training set combines public and internal data sets. The public data is composed of all the MT systems submitted to WMT News (or General) Translation task between years 2009 and 2023. Our internal data set is prepared by translating parallel data using four MT systems: Moses SMT (Koehn et al., 2007), readily available bilingual NMT (Tiede-

[1] Metrics and usage instructions are available at: https://github.com/marian-nmt/wmt23-metrics

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 751–755
December 6–7, 2023.

mann and Thottingal, 2020), multilingual transformer NMT (Gowda et al., 2021), and Microsoft Translator service. The number of examples in our training data is reported in Table 1.

For each training example i, let s_i, r_i and h_i, be source, reference and MT hypothesis segments, respectively. Each example is initially scored using teacher metrics that use reference translations and later references are dropped while training the student metrics. In this work, we use `COMET22` (Rei et al., 2022a) and `ChrF` (Popović, 2015) as teacher metrics. Teacher metrics that need source, reference and hypotheses as inputs – e.g. `COMET22` – produce training data in the form of $(s_i, r_i, h_i) \rightarrow \mathbb{R}$. The reference-only teachers such as `ChrF` produce $(r_i, h_i) \rightarrow \mathbb{R}$. All teacher sentence-level scores are normalized to the $[0, 1]$ range. For `COMET22` this required no change; for `ChrF`, computed by SacreBLEU (Post, 2018), we divide scores by 100.

Distilled students are trained on source-hypothesis pairs $(s_i, h_i) \rightarrow \mathbb{R}$ where the score is from the respective original reference-based teacher. Neither the references nor the human scores are directly seen by the student. However, indirectly, human scores may have been used by the teacher metric, which is the case for `COMET22`, but not for `ChrF`.

Dataset	Number of Examples
WMT09-21 systems	4.0M
WMT22 systems	0.5M
WMT23 systems	0.5M
Internal dataset	6.8M

Table 1: Training dataset size.

2.2 Model

Our distilled models have a similar architecture to `COMET-QE` models (Rei et al., 2020a),[2] and are implemented in MarianNMT (Junczys-Dowmunt et al., 2018), a fast NMT toolkit.[3] We slightly simplify the architecture by removing the encoder layer mixing and the batch-normalization present in the original implementation (neither seemed to contribute to any improvements), but we keep the general architecture of the added FFN regressor and the way how the encoder embeddings of source and hypothesis are combined into a single vector. Final output scores are squashed to the $[0, 1]$ range via a sigmoid function.

Similar to `COMET22`, we initialize our student models with the pretrained weights from InfoXLM (Chi et al., 2021),[4] specifically `infoxlm-large` that has 24 transformer layers (Vaswani et al., 2017).

We create the following four student models:

- `Cometoid22-wmt21`: student model distilled from `COMET22` and trained on scored data from the WMT News Translation task from 2009 - 2021 and similarly sized private data.
- `Cometoid22-wmt22`: Same as above, except we include system outputs submitted to WMT22. This is our *primary* submission to WMT23 Metrics shared task.
- `Cometoid22-wmt23`: Same as the above, except we include the system outputs submitted to WMT23.
- `ChrFoid-wmt23`: Same as the above, but we use segment-level `ChrF` as the teacher. This is an experimental model trained after the WMT23 Metrics shared-task deadline and has not been submitted to the shared task.

We evaluate our models on the WMT22 shared task while including WMT22 shared-task system outputs (MT systems and their reference-based scores) in the training data. This may seem suspicious at first, but note that our models do not use any human scores (the actual ground-truth of the task) in the training process, neither did the reference-based teachers which were trained before the WMT22 shared task. For the part of the evaluation where system submissions are available, this can be seen as part of an involved scoring process where the teacher remains blind to WMT22/WMT23 outputs, but the student does see them during distillation.

However, we are aware that this view may be disputable, hence we have submitted our `Cometoid22-wmt22` (blind to WMT23 outputs) as the primary submission to the WMT23 shared task instead of `Cometoid22-wmt23` that has seen scored WMT23 outputs (but not the actual ground-truth). We also provide results for `Cometoid22-wmt21` which is fully blind in regard to both – WMT22 and WMT23 outputs.

[2] https://huggingface.co/Unbabel/wmt20-comet-qe-da

[3] https://marian-nmt.github.io

[4] https://huggingface.co/microsoft/infoxlm-large

Metric	DA+SQM	MQM
Metricx_xxl_MQM_2020	**0.861**	0.850
Metricx_xl_MQM_2020	0.859	0.843
Cometoid22-wmt23 QE	0.859	0.803
Metricx_xxl_DA_2019	0.857	**0.865**
Cometoid22-wmt22 QE	0.857	0.807
Metricx_xl_DA_2019	0.850	**0.865**
Cometoid22-wmt21 QE	0.848	0.788
UniTE	0.847	0.828
COMET22	0.839	0.839
UniTE-ref	0.838	0.818
COMETKiwi(WMT22) QE	0.832	0.788
Cross-QE QE	0.832	0.781
ChrFoid-wmt23 QE	0.832	0.777
COMETKiwi (public) QE	0.816	0.770
ChrF	0.758	0.734

Table 2: WMT22 Evaluation system-level pairwise accuracy with DA+SQM (13 language pairs) and MQM (3 language pairs only). Rows are ordered by DA+SQM accuracy. Cometoid22 metrics are the best reference-free (QE) metrics.

2.3 Training

We ensure that scores from teacher metrics are in $[0, 1]$ range and optimize student metrics using cross-entropy loss.[5] Rei et al. (2020b) found that freezing InfoXLM layers for a number of epochs and training only the added parameters is beneficial, however, we were unable to confirm this with our metrics; we have fine-tuned all parameters till convergence according to perplexity on a small heldout subset of the data. For the final primary submission, we added the heldout data back to the training data and trained for the same number of iterations. We see minor improvements from Mix-up regularization (Pinto et al., 2022) which we use for all student trainings.

3 Results and Analysis

We report system level pairwise accuracy obtained using `mt-metrics-eval`,[6] the official meta-evaluation pipeline used in WMT22 Metrics task. Table 2 shows that our COMETOID metrics are the top-performing QE metrics on the WMT22 Metrics data set. Interestingly, COMETOID student models also outperform the `COMET22` reference-based teacher model on DA+SQM data (we do fare worse on the smaller MQM data set only). Last but not least, `ChrFoid` – our student metric distilled from the `ChrF` (Popović, 2015) string-based metric – does surprisingly well and out-performs the teacher metric by a considerable margin despite now being reference-free.

[5] Our preliminary experiments with mean absolute error loss performed inferior to cross-entropy.

[6] https://github.com/google-research/mt-metrics-eval

4 Related Work

Reference-free (QE) metrics: Comet20-QE (Rei et al., 2020b) and CometKiwi22 (Rei et al., 2022c) are popular QE metrics. UniTE (Wan et al., 2022) supports inference in reference-free mode, in addition to reference-based mode. These metrics rely on scores from human evaluators during training and are limited by availability of high quality human ratings. Our metrics are trained with scores from teacher models and are trained on larger training data than what has been rated by human evaluators.

Distillation: Pu et al. (2021) and Rei et al. (2022b) apply knowledge distillation to the reference-based metrics, however, their distillation is aimed at reducing the model size for the sake of reducing computational cost during inference. Our work differs from theirs, as we distill with the aim of removing the need for human references at inference time.

5 Conclusion

We believe this work describes a perhaps simpler avenue towards more powerful QE metrics than proposed so far: *build strong reference-based first, next distill into even stronger QE metrics*. It further seems that performance improves with adding fully synthetic data (via adding larger amounts of inputs and automatically scored outputs). This effect seems also applicable to "dumb" metrics like `ChrF`: we have arrived at CHRFOID, a QE metric that has seen no human scores at all, and yet rivals the performance of the best previously available QE metrics. Knowledge distillation combined with a strong information bottleneck (reference-based to reference-free) seems to be the key in this new approach.

Limitations

Using available system outputs of the *same* shared task for training the metric may be a disputable approach even if the ground-truth was not used. Training time and model size of our distilled metrics are similar to the other popular metrics, and may be a limitation.

Ethics Statement

Knowledge distillation of existing models is always close to "model-stealing". The information provided here should be used responsibly and with publicly available models or according to terms of service.

Acknowledgements

Authors like to thank Roman Grundkiewicz for sharing some of the private data sets used in this work; Hieu Hoang for help with training Moses SMT, one of many MT systems used to create training data for distillation. Authors also like to thank the developers of `mt-metrics-eval` for creating and open-sourcing the meta-evaluation tool.

References

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Thamme Gowda, Zhao Zhang, Chris Mattmann, and Jonathan May. 2021. Many-to-English machine translation tools, data, and pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 306–316, Online. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Young Jin Kim, Marcin Junczys-Dowmunt, Hany Hassan, Alham Fikri Aji, Kenneth Heafield, Roman Grundkiewicz, and Nikolay Bogoychev. 2019. From research to production and back: Ludicrously fast neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 280–288, Hong Kong. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Percy Liang, Hal Daumé, and Dan Klein. 2008. Structure compilation: Trading structure for features. ICML '08, page 592–599, New York, NY, USA. Association for Computing Machinery.

Francesco Pinto, Harry Yang, Ser Nam Lim, Philip Torr, and Puneet Dokania. 2022. Using mixup as a regularizer can surprisingly improve accuracy & out-of-distribution robustness. In *Advances in Neural Information Processing Systems*, volume 35, pages 14608–14622. Curran Associates, Inc.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*,

pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, José G.C. de Souza, Pedro G. Ramos, André F.T. Martins, Luisa Coheur, and Alon Lavie. 2022b. Searching for COMETINHO: The little metric that could. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 61–70, Ghent, Belgium. European Association for Machine Translation.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020a. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020b. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022c. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 2022. UniTE: Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127, Dublin, Ireland. Association for Computational Linguistics.

Chunting Zhou, Jiatao Gu, and Graham Neubig. 2020. Understanding knowledge distillation in non-autoregressive machine translation. In *International Conference on Learning Representations*.

MetricX-23: The Google Submission to the WMT 2023 Metrics Shared Task

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and **Markus Freitag**
Google
{jjuraska,marafin,dandeutsch,adisid,mahdim,freitag}@google.com

Abstract

This report details the MetricX-23 submission to the WMT23 Metrics Shared Task and provides an overview of the experiments that informed which metrics were submitted. Our 3 submissions—each with a quality estimation (or reference-free) version—are all learned regression-based metrics that vary in the data used for training and which pretrained language model was used for initialization. We report results related to understanding (1) which supervised training data to use, (2) the impact of how the training labels are normalized, (3) the amount of synthetic training data to use, (4) how metric performance is related to model size, and (5) the effect of initializing the metrics with different pretrained language models. The most successful training recipe for MetricX employs two-stage fine-tuning on DA and MQM ratings, and includes synthetic training data. Finally, one important takeaway from our extensive experiments is that optimizing for both segment- and system-level performance at the same time is a challenging task.[1]

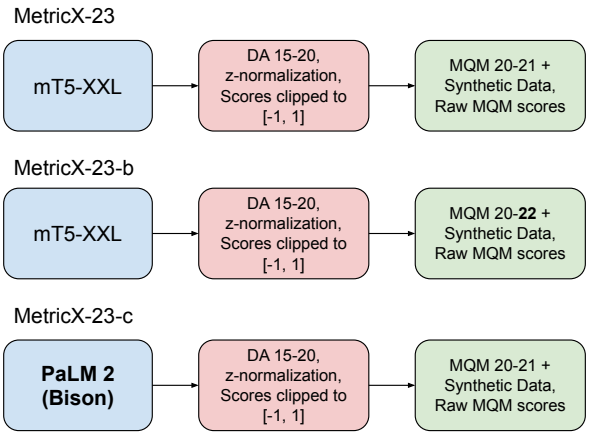


Figure 1: A high-level summary of the 3 different submissions to the WMT'23 Metrics Shared Task. MetricX-23-b and MetricX-23-c differ from the primary submission in that the "b" version is finetuned on MQM 2022 data in addition to 2020 and 2021, and the "c" version uses PaLM 2 as its pretrained language model (differences in bold). Each of the submission also includes a QE variant that follows the same training procedure.

1 Introduction

Automatic evaluation metrics are critical to the development of machine translation (MT) systems. They are the most frequently used method for comparing two MT systems and deciding which generates higher quality translations. Each year, the Conference on Machine Translation (WMT) runs a Metrics Shared Task to benchmark the quality of state-of-the-art evaluation metrics (Freitag et al., 2022). Meta-evaluating metrics by measuring how well they correlate to human ratings of translation quality is critical for understanding the extent to which automatic evaluations of MT systems are trustworthy.

This report details the MetricX-23 submission to the Metrics Shared Task. MetricX is a learned regression-based metric that is trained to predict a floating point score that represents the quality of a candidate translation. This year, we made 3 different submissions to the shared task that vary in the training data that is used for finetuning and which pretrained language model is used for initialization. Our primary submission, denoted MetricX-23, is based on the mT5 encoder-decoder language model (Xue et al., 2021), which is further finetuned on direct assessment (DA) ratings, MQM data (Lommel et al., 2014; Freitag et al., 2021), and synthetic data. Our contrasting submission, MetricX-23-b, includes additional MQM data, and MetricX-23-c finetunes the PaLM 2 language model (Anil et al., 2023) instead of mT5. Each of the 3 submissions has a reference-based and quality estimation (QE, or reference-free) version.

Figure 1 contains a high-level overview of the training recipe that we used for our submissions. In

[1]Our code and mT5-based models can be found at https://github.com/google-research/metricx.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 756–767
December 6–7, 2023.

order to arrive at the metrics that were ultimately submitted to the shared task, we ran various experiments that are detailed in this report. The key takeaways from those experiments include:

1. Training on z-normalized DA scores instead of the raw scores tends to be a trade-off between segment- and system-level performance;
2. Training on raw MQM ratings is better than z-normalized ratings;
3. Training on DA data followed by MQM data yields a better metric than on either type of data individually;
4. Synthetic data is necessary for the metric to learn to score the reference against itself higher than a machine translation against the reference;
5. Metric performance improves significantly as the size of the pretrained language model increases.

2 Metric Descriptions

The MetricX-23 metrics that were submitted to the Metrics Shared Task are all learned regression-based metrics that are trained to predict a floating point number that represents the quality of a given translation.

The input to the reference-based metrics is the candidate translation (hypothesis) and reference segments—each with a corresponding prefix ("candidate:" and "reference:", respectively)—concatenated together. The combined input is encoded by the model, and then the metric uses the encoding to predict a score. This stands in contrast to COMET-style metrics in which the hypothesis and reference are encoded separately, then combined in order to predict a score (Rei et al., 2020). The QE variants use the source segment instead of the reference, with the prefix changed to "source:".

We use two different network architectures for different versions of the metric. The choice of architecture depends on which pretrained language model is used to initialize the model.

The first architecture is based on the encoder-decoder mT5 language model (Xue et al., 2021). The input is encoded by the encoder, then the output logit from an arbitrary token in the vocabulary distribution from the first step of decoding is selected to represent the score for the hypothesis and trained accordingly.[2] In practice, we found that this method for using the pretrained weights for both the encoder and decoder worked better than using a regression head on top of the encoder and discarding the decoder.

The second architecture is the prefix language model based on Transformer (Vaswani et al., 2017) used by the PaLM 2 model (Anil et al., 2023). We augment the architecture by adding a feedforward regression layer on top of the input encoding. The output from the feedforward layer is trained to predict the translation quality score.

Both types of model are trained with a mean squared error (MSE) loss function. Further implementation details related to checkpoint selection, optimization, etc., can be found in §3.3. Information related to training data, label normalization, etc., can be found in §4.

3 Experimental Setup

3.1 Training and Evaluation Data

The two data sources that are primarily used to train and meta-evaluate MT metrics are the direct assessment (DA) data and Multi-dimensional Quality Metrics (MQM; Lommel et al., 2014; Freitag et al., 2021) that have been collected by WMT over the years, and both of which are publicly available. We use the DA data for training and the MQM data for both training and evaluation.

The DA judgments come from non-expert annotators that score the quality of a translation on a scale from 0 to 100. Often, the scores are z-normalized per rater in order to better compare across raters, since each rater may have a different rating strategy despite using the same scale. We experiment with using different subsets of the DA data from 2015 to 2021, as well as using the raw rating or z-normalized rating as the ground-truth quality score.

In contrast, the MQM ratings are done by professional raters. Each rater marks specific spans of text within a translation that contain an error, and label that error with a severity level and category. Each error receives a weight based on its severity and category. A segment's MQM score is the sum of the error weights in the segment. Our experiments use the MQM data collected by WMT

[2]The specific token to use can be chosen arbitrarily from the vocabulary, but the same token is then used throughout the training and inference. In our case, we opted for one of the `<extra_id_**>` tokens reserved in mT5's vocabulary.

from 2020 to 2022. The 2022 ratings are our primary evaluation dataset, and all correlations that we report are calculated on this dataset.

Our experiments additionally leverage a metrics challenge set, DEMETR (Karpinska et al., 2022), for metric meta-evaluation. DEMETR is a collection of paired translations that probe a metric's ability to correctly model different phenomena. The pairs of translations differ by some linguistic phenomena, with one translation assumed to be higher-quality than the other, and the metric is evaluated on how often it correctly ranks the two translations. We use DEMETR to evaluate how frequently the reference translations evaluated against itself receives a higher score than a machine translation evaluated against the same reference.

3.2 Meta-Evaluation

In our experiments, we calculate the metrics' agreements with human judgments of translation quality using four different correlations. At the system-level, we use both Pearson's r and pairwise accuracy (Kocmi et al., 2021). System-level Pearson's r captures how strong the linear relationship is between the metric and human scores for MT systems. Pairwise accuracy evaluates a metric's ranking of MT systems by calculating the proportion of all possible pairs of MT systems that are ranked the same by the metric and human scores.

At the segment-level, we use the no-grouping Pearson's r and the group-by-item pairwise accuracy with tie calibration as described by Deutsch et al. (2023).[3] The no-grouping Pearson's r quantifies the linear relationship between the metric and human scores across all possible translations from every system and document. The group-by-item pairwise accuracy calculates the proportion of all possible pairs of translations for the same input segment that are ranked the same or predicted to be a tie by the metric and human, then averages the accuracies over all possible input segments. Since regression-based metrics rarely predict ties and the segment-level pairwise accuracy rewards correct tie predictions, Deutsch et al. (2023) uses a procedure called tie calibration that automatically introduces ties into metric scores by introducing an ϵ such that any two translations with a difference in metric score less than ϵ are considered to be tied.

[3]We chose this pairwise accuracy over Kendall's τ, which has typically been used in WMT Metrics evaluation, for its superior handling of ties in metric scores.

3.3 Implementation Details

Our metrics are implemented with TensorFlow (Abadi et al., 2015) and the T5X library (Roberts et al., 2022). Each training run uses 64 TPUs and trains for a maximum of 10K steps with a batch size of 512 on the DA data, or 3K steps with a batch size of 256 on the much smaller MQM dataset. Adafactor is used for optimization (Shazeer and Stern, 2018). Checkpoint selection is done by selecting the model that has the highest segment-level pairwise accuracy after tie calibration on the en-de and zh-en language pairs.

We are publicly releasing our mT5-based submissions, converted from TensorFlow to PyTorch (Paszke et al., 2019) checkpoints, along with corresponding code to use them to predict translation quality scores.

4 Experimental Results

We made three different submissions to the shared task, each with a reference-based and QE variant:

1. **MetricX-23(-QE):** An mT5-XXL model that was finetuned on a combination of DA data from 2015–2020, MQM data from 2020–2021, and synthetic data.
2. **MetricX-23(-QE)-b:** The same as MetricX-23(-QE) except we additionally included MQM data from 2022.
3. **MetricX-23(-QE)-c:** The same as MetricX-23(-QE) except it is a finetuned PaLM-2 Bison model.

An overview of these submissions is shown in Figure 1. In the rest of this section, we describe the experimental results that led us to these submissions.

The experiments in the process of determining the best training recipe were performed with mT5-XL (3.7B parameters), but our final submissions then use the XXL variant with 13B parameters. All results are reported as the mean of 3 independent runs, along with the standard deviation across the runs, unless stated otherwise.

4.1 Training Data

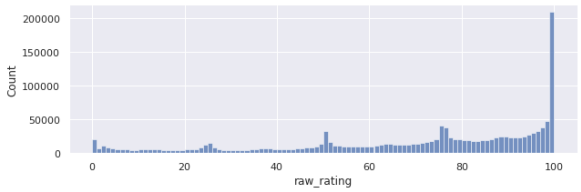
We start by determining which of the data available from previous WMT Metrics Shared Tasks is useful for training our metric, and whether it is beneficial to perform any transformations of the ratings before using them for training.

4.1.1 DA Ratings

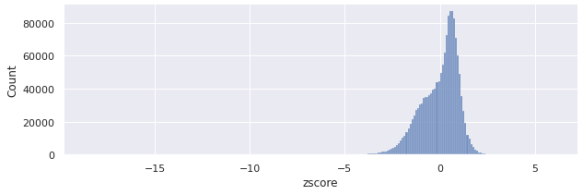
DA ratings have been used for scoring candidate translations in the shared task since 2015. To obtain DA ratings, human annotators were asked to provide an integer score on a scale of 0 to 100 for a translation produced by an MT system, given a reference translation produced by an expert translator.[4] There are over 2M raw DA ratings available from the years between 2015 and 2021, spanning 40 different language pairs. The total number of ratings drops to ca. 1M when ratings for the same segment (from different raters) are aggregated, which we do in order to avoid providing the model with different signals for the same translation.

Since the DA ratings come from hundreds of different raters, the WMT Metrics Shared Task organizers typically z-normalize them per rater before using them as the ground truth for metric evaluation. This is to make the ratings more comparable across different raters, considering some of them can be very strict, others lenient, and some can use the whole rating scale, while others just a narrow range of it. Hence a DA rating of, say, 50 can end up being used for translations of widely varying quality. The normalization ensures that the mean of each rater's ratings is 0 and the standard deviation is 1. These official normalized DA ratings, which we refer to as z-scores, are available along with the *raw* DA ratings in the data from the shared tasks.

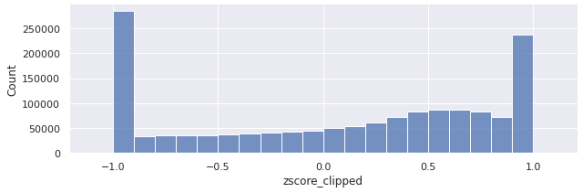
Score Normalization. In our first experiment, we compare the performance of our metric finetuned on the raw DA ratings and on the z-scores on different subsets of the DA data. As the first two rows of Table 1 show, using z-scores results in an overall weaker performance, but a drastic improvement on segment-level Pearson's r (42.93 vs. 38.83 for zh-en). Thus, picking between using raw ratings or z-scores for training the metric comes down to the preference between high *system*-level or high *segment*-level performance. Given the models' performance on the system-level metrics is already relatively high (between 80 and 99), we choose to use z-normalized DA ratings over their raw counterparts for our submissions. Nevertheless, as we show in Section 4.2, adding synthetic data during finetuning can restore some of the system-level performance. Moreover, our later experiments in Section 4.3 demonstrate that further finetuning on MQM ratings is more effective using a model first finetuned on DA z-scores than raw DA ratings.



(a) Raw DA ratings.



(b) DA z-scores.



(c) Clipped DA z-scores.

Figure 2: WMT15-20 DA rating distributions after z-normalization and after clipping to $[-1.0, 1.0]$, compared to the raw rating distribution.

Score Clipping. Normalized and raw ratings follow very different distributions, as depicted in Figure 2. The raw rating distribution is relatively flat across the whole range with a large spike at the maximum value, i.e., 100. In contrast, the z-score distribution ranges roughly from -17 to 5, with the majority of the mass between -1 and 1, a sharp peak around 0.65, and a long tail on the negative side. In order to prevent the model from putting too much weight on the outliers during training, and not learning to differentiate well between translations scored around zero, we propose clipping the scores to be in the $[-1.0, 1.0]$ range.[5] This creates a spike on the right end, similar to that observed on the raw rating distribution, but also a spike on the left end, similar in magnitude to the other spike (see Figure 2c). Finetuning a model on the clipped z-scores results in segment-level performance gains compared to the unmodi-

[4] Technically, some of the translation data is "target-original", meaning that the reference (target) is the text originally to be translated, and the source is the translation. This is the case for some language pairs in the DA data from earlier years.

[5] MSE loss magnifies errors in predictions greater than 1 and shrinks errors smaller than 1 relative to the absolute difference.

	SEG pairwise acc.		SEG Pearson		SYS pairwise acc.		SYS Pearson	
	en-de	zh-en	en-de	zh-en	en-de	zh-en	en-de	zh-en
15-20 raw	**59.98** ±0.03	**51.63** ±0.15	43.51 ±0.77	38.83 ±0.44	**83.33** ±1.28	**89.38** ±1.68	**90.79** ±1.39	**98.51** ±0.35
15-20 z	59.57 ±0.20	51.26 ±0.40	47.08 ±0.23	42.93 ±0.13	81.62 ±1.48	85.35 ±1.68	85.60 ±1.00	98.02 ±0.20
15-20 z clipped	59.77 ±0.21	51.45 ±0.15	46.89 ±0.32	**45.05** ±0.24	79.49 ±0.00	86.08 ±1.68	83.24 ±0.97	97.28 ±0.11
15-21 z clipped	59.70 ±0.10	50.76 ±0.30	**47.76** ±1.02	43.26 ±0.70	80.34 ±1.96	85.35 ±1.68	85.21 ±1.67	97.31 ±0.82

Table 1: Performance of MetricX that is initialized with mT5-XL, finetuned on DA ratings in different ways. "15-20" indicates that ratings from the years 2015 through 2020 were used, "z" indicates z-normalized scores, and "clipped" denotes experiments with the scores clipped to the $[-1.0, 1.0]$ range. Note that the data from 2015 is a small set of 2,500 z-normalized ratings only, so there is technically no data from 2015 in the "15-20 raw" setting.

fied z-scores, though further sacrifices some of the system-level performance (compare rows 2 and 3 in Table 1). Compared to finetuning on raw ratings (row 1 vs. row 3), there is up to 16% increase on the segment level (Pearson on zh-en), at the cost of an up to 9% drop in system-level performance (Pearson on en-de).

Data Selection. The DA ratings we used in our experiments thus far were from 2015 to 2020. Saving the WMT22 data for the validation set, we have the option of adding the WMT21 DA ratings to the training set. Doing this leads to moderate gains in system-level en-de performance, yet an equal, if not bigger, performance drop in zh-en across all metrics (compare the last two rows in Table 1). We also tried excluding earlier years of DA ratings, such as 2015–2017 or 2015–2018, since up until 2018 a half of the translations in the data were target-original (Barrault et al., 2019), and all DA annotations were reference-based (Ma et al., 2019), as opposed to source-based, such as is the case with a good part of the DA annotations from 2019 onward, and all of the MQM ratings. We hypothesized that the older data might thus be providing some low-quality signals to the model during training, negatively affecting the performance. Nevertheless, the model seems to prefer additional training data, even if of a mixed quality, as we consistently observed a slight drop in performance after excluding the earlier years, especially on system level. Therefore, the rest of the experiments uses DA data from 2015 to 2020.

4.1.2 MQM Ratings

MQM ratings have been collected in the context of the WMT Metrics Shared Task only since 2020. Due to the MQM annotation being significantly more labor-intensive, there is significantly less data collected per year than using the DA methodology. In fact, MQM ratings are only available for three language pairs, namely en-de, zh-en and en-ru. Since we reserve the ratings from 2022 for validation, we are left with only two years worth of MQM data to use for training, which amounts to approximately 114K ratings. For both training and evaluation we negate the MQM scores, changing thus the range to $[-25, 0]$, so that the score corresponding to no errors in the translation would be the highest value, as opposed to the lowest value, in the range. In the following paragraphs, we discuss our experiments with finetuning a model on MQM ratings only, in order to see if the model learns anything different than what it learns from the DA ratings.

Data Selection. We start this set of experiments with finetuning our model on the combination of '20 and '21 MQM ratings, and confirming that there is an added benefit to it over finetuning on just one of the years. As demonstrated by the first two rows in Table 2, finetuning on the '20 and '21 data individually leads to very different performance across the set of metrics. Using just the '20 MQM data by itself, our model achieves better segment-level performance than using all of the DA data, however, it is the opposite case on system level. As for training on '21 data only, the performance is significantly worse overall despite a slightly better segment-wise pairwise accuracy on en-de. Although this may suggest that the '21 MQM data is of a lower quality, it may also simply be the consequence of the '21 data having only a little over a third of the number

	SEG pairwise acc.		SEG Pearson		SYS pairwise acc.		SYS Pearson	
	en-de	zh-en	en-de	zh-en	en-de	zh-en	en-de	zh-en
20 raw	59.91 ±0.24	52.52 ±0.13	**48.03** ±1.56	54.43 ±0.58	**77.35** ±2.67	**84.98** ±0.63	**75.30** ±0.71	85.46 ±1.76
21 raw	60.43 ±0.12	51.92 ±0.04	42.21 ±1.69	54.07 ±1.00	69.66 ±1.96	82.42 ±1.10	56.62 ±4.67	85.92 ±0.40
20-21 raw	**60.77** ±0.17	**52.72** ±0.24	47.59 ±1.82	**55.41** ±0.43	73.93 ±0.74	83.52 ±1.10	72.70 ±0.15	85.68 ±0.70
20-21 z	59.74 ±0.33	51.98 ±0.82	45.84 ±0.99	54.34 ±0.75	74.79 ±0.74	82.78 ±2.29	70.79 ±2.79	**86.14** ±1.06

Table 2: Performance of MetricX initialized with mT5-XL on the WMT22 MQM dataset, finetuned on different subsets of MQM ratings. "z" indicates z-normalized scores.

of ratings in the '20 data.[6] Finetuning on both the '20 and the '21 data combined, however, outperforms both of the individual years on segment-level metrics, while it lands somewhere in between according to system-level metrics (see row 3 in the table). Hence, we use the MQM ratings from both years in all of our subsequent experiments.

Score Normalization. We observed on the DA data that z-normalization has certain benefits, so we experiment with it even on the MQM data. We perform the z-normalization ourselves in the same way the shared task organizers normally do for the DA ratings, and compare a model finetuned on these z-scores to one finetuned on the raw MQM ratings. It is clear from the comparison of rows 3 and 4 in Table 2 that z-normalization drags the performance down, especially on the segment level.[7] One possible explanation could be that the normalization has a negative effect here because of the raters having annotated different sets of documents each, and the set of raters being very small at the same time. It could also be that z-normalization is actually not a very practical transformation of training labels for this task, yet it helps in case of DA ratings, which are of a much lower quality.[8] The metrics are evaluated against raw MQM ratings, so z-normalization during training could negatively impact its Pearson correlation at test time. At any rate, based on this result, we opt for the raw MQM ratings when finetuning our models henceforth.

[6] In the '20 MQM data, candidate translations have multiple ratings, so we also experimented with averaging them, but that, somewhat surprisingly, resulted in a consistently lower performance across the board.

[7] We verified that this is the case both with and without aggregating the ratings of the same translations by multiple raters in the '20 data.

[8] For instance, z-normalization discounts the ratings of raters who gave the same score, e.g., 100, to most of the translations they rated.

4.2 Adding Synthetic Data

Using the DEMETR challenge set, we discovered that training MetricX on either the DA or the MQM dataset does not teach it to reliably score a translation that exactly matches the reference higher than or equal to a machine translation, which should be a basic sanity check for an evaluation metric. In order to fix this behavior, we create simple synthetic examples where the reference is copied as the candidate translation and the label is set to the maximum score. Depending on the training set these synthetic examples are used along with, the labels may need to be rescaled to ensure they correspond to the maximum score, e.g., 100 when used with raw DA ratings or 0 with MQM ratings. Since z-scores do not have a maximum value, per se, there is no straightforward way of incorporating this synthetic data into such a training set. Clipped z-scores make this trivial though, which is another argument for training MetricX on clipped DA z-scores instead of the full range of z-scores. We use all of the references across all language pairs in the DA data between 2015 and 2022 to construct this synthetic dataset, which amounts to a little over 180K unique examples.

We also prepare a second synthetic dataset with the opposite type of examples, that is, ones that have no candidate translation and therefore a label corresponding to the minimum score (i.e., 0 for DA, and -25 for MQM). It is created using the same data as the other synthetic dataset, only instead of copying the reference, we set all of the translations to an empty string, resulting in the same number of synthetic examples.

Next, we perform experiments to determine what the minimum ratio of synthetic examples to regular training examples is with which a high accuracy on

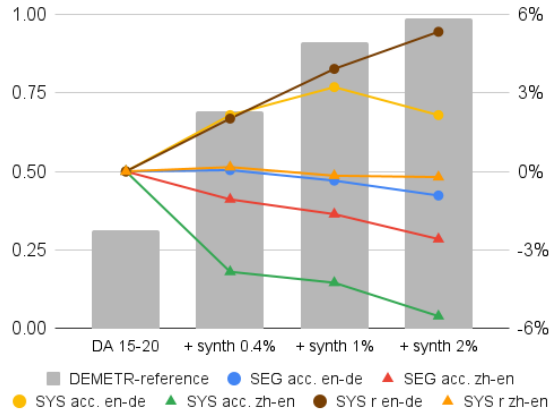


Figure 3: Effects of adding different proportions of synthetic data to the DA 15-20 training set (using clipped z-scores). The gray bars indicate the DEMETR-reference score (with the scale on the left y-axis), while the lines show the relative increase or decrease of the correlation metrics w.r.t. using no synthetic data (with the scale on the right y-axis). "Acc." stands for pairwise accuracy, and "r" denotes Pearson's r. Segment-level Pearson's r followed similar trends to segment-level accuracy, and was therefore omitted for better readability.

the DEMETR-reference metric can be achieved.[9] Figure 3 illustrates the effects of adding the synthetic data to the "DA 15-20 z clipped" training set in the following proportions: 0.4%, 1%, and 2% of the DA examples.[10] As we can see by looking at the gray bars, the synthetic data has the desired effect of bringing the accuracy of correctly scoring a reference higher than a machine translation from a mere 31% to almost 100%. However, this comes at a cost. With the increasing amount of synthetic data, the segment-level performance slowly degrades for both language pairs (see the blue and red lines in the plot). On the system level side, we see mixed results: a big improvement for en-de (see the yellow and brown lines) and a comparable drop in performance for zh-en, though only according to pairwise accuracy (see the orange and green lines). As such, we find adding 1% of synthetic examples to the DA data a good compromise between achieving relatively high DEMETR scores and maintaining good performance on the segment- and system-level correlation metrics. For the actual scores, we refer the reader to row 1 in Table 3.

We carried out a similar study with the MQM 20-21 training data, landing on 2% of synthetic examples as the best proportion.[11] Note that, with a model finetuned on raw MQM ratings, the accuracy on DEMETR was relatively high to begin with, typically between 80% and 90%. The 2% of synthetic data brought it up to nearly 99% though. Interestingly, training on the combination of MQM and synthetic data did not have a negative impact on the correlation metrics, as was the case with DA data. On the contrary, the performance received a consistent boost across all metrics, with an up to 1% increase in system-level scores and 5% in segment-level scores (see row 2 in Table 3).

4.3 Two-Stage Finetuning

In the previous two sections, we identified the best subset and format of the DA/MQM ratings, and we found the right balance between the DA/MQM examples and the synthetic examples for training a MetricX model. Here, we take it one step further and perform two-stage finetuning experiments, wherein we first finetune the model on the DA 15-20 training set, and then further finetune it on MQM 20-21 data with a smaller learning rate. The reason for finetuning on the two datasets in this order is 3-fold: (1) MQM is substantially smaller and has a limited language coverage, (2) MQM is a higher-quality dataset, and (3) the metric's performance is ultimately evaluated on MQM ratings (whether it be our validation set, or the shared task's test set).

Of all the four combinations of raw and z-normalized DA and MQM ratings, we found, somewhat surprisingly, that using DA z-scores (aggregated) in the first stage followed by raw MQM ratings in the second stage leads to the best results. Intuitively, using raw ratings in both stages, or z-scores in both stages, should provide a smoother learning process for the model, as it does not need to relearn the label scales.[12] Nevertheless, it appears that the neural model is not negatively affected by the shift, and instead it prefers learning from the DA and MQM data in their respective most effective format.

Before diving into the two-stage experiment results, let us recap that, so far, MetricX achieved the best segment-level performance when finetuned on the MQM 20-21 dataset, and the best system-level performance on the DA 15-20 dataset (in fact, substantially better than on the MQM dataset). The top two rows in Table 3 show the performance of models finetuned on these datasets individually,

[9] When included in the training data, both of the synthetic datasets are added in the same proportion.

[10] 1% corresponds to ca. 10K synthetic examples.

[11] Since the MQM dataset is much smaller, 2% here corresponds to only ca. 2K synthetic examples.

[12] In the experiments with raw ratings in both stages, we rescaled the DA ratings to the $[-25, 0]$ range, so as to match the MQM scale used then in the second stage.

	SEG pairwise acc.		SEG Pearson		SYS pairwise acc.		SYS Pearson		DEMETR
	en-de	zh-en	en-de	zh-en	en-de	zh-en	en-de	zh-en	
DA_{syn}	59.56	50.61	46.75	43.44	**82.05**	82.42	**86.50**	**97.12**	91.03
MQM_{syn}	60.96	52.93	49.98	56.12	73.08	83.88	73.49	85.72	**98.57**
DA $\rightarrow$ MQM	61.50	**54.09**	51.59	58.53	73.93	**86.45**	75.46	88.93	46.20
DA $\rightarrow$ MQM_{syn}	**61.61**	54.05	51.55	58.81	75.21	85.71	76.65	89.89	91.93
DA_{syn} $\rightarrow$ MQM	61.46	53.70	52.20	57.97	73.93	85.71	75.63	88.39	82.03
DA_{syn} $\rightarrow$ MQM_{syn}	61.35	53.61	52.30	57.57	75.64	86.08	77.18	89.35	97.53
DA^{21}_{syn} $\rightarrow$ MQM	61.49	53.67	51.36	58.07	74.79	84.62	76.30	89.26	92.30
DA^{21}_{syn} $\rightarrow$ MQM_{syn}	61.45	53.93	**53.62**	**58.88**	75.21	85.71	78.15	89.85	98.17

Table 3: Performance of MetricX initialized with mT5-XL, finetuned first on DA 15-20 clipped z-scores, and subsequently on MQM 20-21 raw ratings, with synthetic data added at different stages. DA_{syn} denotes the DA dataset with 1% of synthetic examples, and MQM_{syn} is the MQM dataset with 2% of synthetic examples. For comparison, the first two rows show the performance of models finetuned on DA_{syn} and MQM_{syn} individually. The last two rows correspond to models finetuned in the first stage on the DA dataset with the '21 ratings added but with all into-English language pairs excluded. Standard deviations are omitted in this table for better readability.

with synthetic data included, serving thus as baselines for the following experiments. Now, the third row shows the scores for a model finetuned in the two-stage fashion without any synthetic data. Comparing these results with those of a model finetuned on the MQM dataset only (see row 3 in Table 2), we see a dramatic improvement in performance across the board. For example, the segment-wise accuracy for zh-en increases from 52.72 to 54.09, and Pearson's r for en-de from 47.59 to 52.30. In system-level metrics we see similar gains, such as the accuracy going up from 83.52 to 86.45, and Pearson's r from 72.70 to 75.46 for zh-en. Similarly, rows 2 and 4 in Table 3 can be compared to see a similar difference, only this time for models trained with synthetic data too. This demonstrates a clear benefit of training our MetricX model on both the DA and the MQM data over training it on either of them individually. That being said, the system-level performance still lags significantly behind models finetuned on DA data only, so there appears to be a trade-off between segment- and system-level performance.

Next, we examine whether there is a difference between including synthetic data in the first stage or the second stage of finetuning. Rows 4 and 5 in Table 3 correspond to these two experiments. The scores show that using synthetic examples in the second stage not only ensures a higher DEMETR accuracy (91.93 vs. 82.03), but also higher correlations with the human scores according to virtually all of the other metrics. Moreover, compared to not using synthetic data at all (see row 3 in the table), combining it with the MQM data in the second stage does not generally sacrifice the overall performance, not to mention it almost doubles the DEMETR accuracy.

Finally, we also tried including synthetic data in both finetuning stages, but not until after the shared task's submission window has passed, hence, we did not use this method in our final MetricX version. The gains over using the synthetic data in the second stage only are rather inconsistent, nevertheless the DEMETR score further increases to 97.53 (see row 6).

4.4 Additional Experiments

Although in Section 4.1 we concluded that adding DA ratings from WMT21 to the training set dragged the performance down, we noted that that was the case for the zh-en language pair only. Revisiting this experiment after the submission, we found that excluding the into-English DA data from WMT21 and including the synthetic data as described in Section 4.2 is, in fact, a potentially better training set than the same with the WMT21 ratings omitted altogether. As we can see in Table 3, the model performs consistently better across most of the metrics when finetuned with the WMT21 out-of-English language pairs included (compare rows 7–8 with rows 5–6). So, while this seems to be the best training recipe, it is not the one we followed for the shared task submissions. Instead, we used DA 15-20 (without synthetic data) in the first stage, corresponding to row 4 in the table.

4.5 Scaling Analysis

Having arrived at our final MetricX training recipe using mT5-XL (3.7B parameters) as the pretrained model, we now briefly compare its performance

	SEG pairwise acc.		SEG Pearson		SYS pairwise acc.		SYS Pearson	
	en-de	zh-en	en-de	zh-en	en-de	zh-en	en-de	zh-en
mT5-L	59.63 ±0.34	53.98 ±0.06	50.82 ±2.16	55.91 ±0.53	76.07 ±0.74	**87.55** ±0.63	79.92 ±1.20	**93.63** ±2.48
mT5-XL	61.61 ±0.12	54.05 ±0.29	51.55 ±1.06	58.81 ±0.12	75.21 ±0.74	85.71 ±1.10	76.65 ±1.44	89.89 ±0.23
mT5-XXL	**61.92** ±0.19	**54.76** ±0.13	**54.57** ±0.36	**60.06** ±0.50	**82.48** ±0.74	86.81 ±1.10	**84.49** ±0.30	90.85 ±0.31

Table 4: Performance of MetricX with different variants of mT5 as the initialization model, trained using our final recipe, which involves first finetuning on DA 15-20 ratings and then on MQM 20-21 ratings with synthetic data.

with two other variants of mT5: one smaller (mT5-Large with 1.2B parameters) and one larger (mT5-XXL with 13B parameters). From Table 4 it is clear that MetricX can benefit from a bigger pretrained model for initialization, as mT5-XXL has a good margin on mT5-XL across all metrics. The XXL variant thus becomes our choice for all of our mT5-based submissions to the shared task.

Interestingly, mT5-Large outperforms the two bigger variants in system-level metrics on the zh-en language pair, and that not by a negligible margin. Combined with the results in the earlier sections and our observation that the system-level performance of MetricX is typically highest right at the beginning of training, and quickly declines as the model gradually improves on segment-level metrics, it appears it may be challenging to come up with a single MT evaluation metric that excels at both the segment and the system level. This phenomenon can also be observed among several of the top metrics in the WMT22 Metrics Shared Task (Freitag et al., 2022), as well as in recent large language model-based approaches to automatic MT evaluation (Kocmi and Federmann, 2023).

4.6 Submission Summary

Throughout the whole of Section 4 thus far, we were reporting results averaged across three independent runs, so as to more reliably develop the best training recipe for the reference-based version of MetricX. Here, we present the performance of our individual final submissions to the WMT23 Metrics Shared Task, described at the beginning of this section, including our QE (or reference-free) metric submissions. All of our submissions follow the same recipe—i.e., are first finetuned on the DA 15-20 aggregated and clipped z-scores, and then further finetuned on MQM 20-21 ratings combined with synthetic examples—but differ in (1) the pretrained model used for initialization, (2) whether they use reference or source segments in the input (the latter being used for the QE submissions), and (3) whether the second-stage training set includes the '22 MQM ratings or not. For the QE variants, we followed the same training recipe as the reference-based version; we did not do a significant amount of analysis into whether the design choices we made for the reference-based metric were also the correct decisions for the QE version. For the models that do use the '22 data for finetuning (which we otherwise use as the validation set), we do not report any scores. For these two submissions, we picked the model checkpoint based on the equivalent training runs without the '22 ratings.

Our remaining four submissions have their scores summarized in Table 5. Between the two reference-based variants (rows 1–2), the mT5-based MetricX is dominant in segment-level scores, whereas the PaLM-2-based one has a strong lead on the system level. The story is similar for the two QE variants (rows 3–4), only the segment-level score differences are less pronounced. Overall, on the WMT22 MQM validation set, the QE variants are not very far behind their reference-based counterparts in performance.

5 Related Work

For many years, lexical-based metrics like BLEU (Papineni et al., 2002) and ChrF (Popović, 2015) were the standard method for automatically evaluating MT output. However, as it was demonstrated that learned evaluation metrics correlate to human ratings significantly higher than lexical-based metrics (Freitag et al., 2022), the vast majority of recent research on MT evaluation has used learned metrics.

Learned MT metrics, such as BLEURT (Sellam et al., 2020; Pu et al., 2021) and COMET

MetricX variant	Pretrained model	SEG p. acc.		SEG Pearson		SYS p. acc.		SYS Pearson	
		en-de	zh-en	en-de	zh-en	en-de	zh-en	en-de	zh-en
23	mT5-XXL	**62.09**	**54.84**	**54.21**	**60.06**	**82.05**	86.81	84.19	91.20
23-c	PaLM-2-Bison	61.56	54.17	51.62	52.24	76.92	**92.31**	91.41	**98.61**
23-QE	mT5-XXL	60.64	54.04	49.78	56.41	78.21	87.91	81.29	91.32
23-QE-c	PaLM-2-Bison	60.23	53.96	49.28	52.48	**82.05**	91.21	**93.64**	96.90

Table 5: Meta-evaluation scores of our four MetricX submissions that did not include the '22 MQM data in the training set. Note that these scores correspond to single runs, as opposed to all the previous results that were averaged across 3 runs.

(Rei et al., 2020, 2022a), differ largely in their network architecture and the specific tasks they are trained to do. Our metric is most closely related to BLEURT. Like BLEURT, the MetricX network architecture creates a joint encoding of both the hypothesis and reference translations together, in contrast to COMET-style metrics that encode them independently. Our network is trained only to do either reference-based or reference-free (QE) judgments of sentence-level translation quality, whereas some metrics like UniTE (Wan et al., 2022) are trained to both tasks at the same time or CometKiwi (Rei et al., 2022b), which learns to predict both word-level quality scores in addition to an overall sentence-level score. Other learned metrics, such as MaTESe (Perrella et al., 2022), take an alternative approach to regression-based metrics and derive a sentence-level quality score by identifying error spans in translations, like is done in the human evaluations of MQM.

More recent approaches to MT evaluation leverage large language models (LLMs) to do zero-shot scoring by either directly predicting scalar quality scores or phrase-level error tagging (Kocmi and Federmann, 2023; Fernandes et al., 2023). These approaches typically leverage models that are orders of magnitude larger than metrics that are trained specifically for MT evaluation.

In comparison to the MetricX submission to the WMT22 Metrics Shared Task, this year's submission shares the same architecture, but we performed a significantly larger number of experiments to arrive at the final models, which are detailed in this report. We also explore how metric performance changes as a function of the number of parameters, experiment with initializing with different pretrained language models, and include a QE submission.

6 Conclusion

In this report, we presented in detail our approach to training MetricX-23, a regression-based MT evaluation metric. We submitted six versions of MetricX-23 to the WMT23 Metrics Shared Task, including both reference-based and QE variants. Some of our findings are that (1) training on direct assessment (DA) ratings and subsequently on MQM ratings leads to a significantly better performance than training on either of the two datasets alone, (2) z-normalization of DA ratings helps achieve better segment-level performance, but is not useful for high-quality MQM ratings, and (3) adding a small amount of synthetic data to the training set, targeting a challenge set, can also boost the metric's overall performance.

Throughout our experiments, we observed an undesirable tension between segment- and system-level performance, making it challenging to improve our metric in both aspects at the same time. Nevertheless, increasing the size of the model used to pre-initialize the metric appears to be one reliable way to increase the overall performance, at least to a certain point. Future work may benefit from a better understanding of the trade-off between segment- and system-level performance, and whether it would be better to focus on separate metrics for these two types of MT evaluation.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan,

Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. PaLM 2 Technical Report.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties Matter: Modifying Kendall's Tau for Modern Metric Meta-Evaluation.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André F. T. Martins, Graham Neubig, Ankush Garg, Jonathan H. Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chikiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and AndrÃ© F. T. Martins. 2022. Results of WMT22 Metrics Shared Task: Stop Using BLEU - Neural Metrics Are Better and More Robust. In *Proceedings of the Seventh Conference on Machine Translation*, pages 46–68, Abu Dhabi. Association for Computational Linguistics.

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. DEMETR: Diagnosing Evaluation Metrics for Translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9540–9561, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica*, (12):0455–463.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library.

Stefano Perrella, Lorenzo Proietti, Alessandro ScirÃ¨, NiccolÃ² Campolungo, and Roberto Navigli. 2022. Matese: Machine translation evaluation as a sequence tagging problem. In *Proceedings of the Seventh Conference on Machine Translation*, pages 569–577, Abu Dhabi. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Amy Pu, Hyung Won Chung, Ankur P Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of EMNLP*.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, JosÃ© G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and AndrÃ© F. T. Martins. 2022b. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 634–645, Abu Dhabi. Association for Computational Linguistics.

Adam Roberts, Hyung Won Chung, Anselm Levskaya, Gaurav Mishra, James Bradbury, Daniel Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, Curtis Hawthorne, Aitor Lewkowycz, Alex Salcianu, Marc van Zee, Jacob Austin, Sebastian Goodman, Livio Baldini Soares, Haitang Hu, Sasha Tsvyashchenko, Aakanksha Chowdhery, Jasmijn Bastings, Jannis Bulian, Xavier Garcia, Jianmo Ni, Andrew Chen, Kathleen Kenealy, Jonathan H. Clark, Stephan Lee, Dan Garrette, James Lee-Thorp, Colin Raffel, Noam Shazeer, Marvin Ritter, Maarten Bosma, Alexandre Passos, Jeremy Maitin-Shepard, Noah Fiedel, Mark Omernick, Brennan Saeta, Ryan Sepassi, Alexander Spiridonov, Joshua Newlan, and Andrea Gesmundo. 2022. Scaling Up Models and Data with `t5x` and `seqio`. *arXiv preprint arXiv:2203.17189*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yu Wan, Keqin Bao, Dayiheng Liu, Baosong Yang, Derek F. Wong, Lidia S. Chao, Wenqiang Lei, and Jun Xie. 2022. Alibaba-translate china's submission for wmt2022 metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 586–592, Abu Dhabi. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

GEMBA-MQM: Detecting Translation Quality Error Spans with GPT-4

Tom Kocmi and **Christian Federmann**
Microsoft, One Microsoft Way, Redmond, WA-98052, USA
{tomkocmi,chrife}@microsoft.com

Abstract

This paper introduces GEMBA-MQM, a GPT-based evaluation metric designed to detect translation quality errors, specifically for the quality estimation setting without the need for human reference translations. Based on the power of large language models (LLM), GEMBA-MQM employs a fixed three-shot prompting technique, querying the GPT-4 model to mark error quality spans. Compared to previous works, our method has language-agnostic prompts, thus avoiding the need for manual prompt preparation for new languages.

While preliminary results indicate that GEMBA-MQM achieves state-of-the-art accuracy for system ranking, we advise caution when using it in academic works to demonstrate improvements over other methods due to its dependence on the proprietary, black-box GPT model.

Metric	Acc.	Meta
GEMBA-MQM	96.5% (1)	0.802 (3)
XCOMET-Ensemble	95.2% (1)	0.825 (1)
docWMT22CometDA	93.7% (2)	0.768 (9)
docWMT22CometKiwiDA	93.7% (2)	0.767 (9)
XCOMET-QE-Ensemble	93.5% (2)	0.808 (2)
COMET	93.5% (2)	0.779 (6)
MetricX-23	93.4% (3)	0.808 (2)
CometKiwi	93.2% (3)	0.782 (5)
Calibri-COMET22	93.1% (3)	0.767 (10)
BLEURT-20	93.0% (4)	0.776 (7)
MaTESe	92.8% (4)	0.782 (5)
mre-score-labse-regular	92.7% (4)	0.743 (13)
mbr-bleurtxv1p-qe	92.5% (4)	0.788 (4)
KG-BERTScore	92.5% (5)	0.774 (7)
MetricX-23-QE	92.0% (5)	0.800 (3)
BERTscore	90.2% (7)	0.742 (13)
MS-COMET-QE-22	90.1% (8)	0.744 (12)
embed_llama	87.3% (10)	0.701 (16)
f200spBLEU	86.8% (11)	0.704 (15)
BLEU	85.9% (12)	0.696 (16)
chrF	85.2% (12)	0.694 (17)

Table 1: Preliminary results of the WMT 2023 Metric Shared task. The first column shows the system-level accuracy, and the second column is the Metrics 2023 meta evaluation. Metrics with gray background need human references. The table does not contain the worst-performing, non-standard metrics due to space reasons.

1 Introduction

GEMBA-MQM builds on the recent finding that large language models (LLMs) can be prompted to assess the quality of machine translation (Kocmi and Federmann, 2023a). We release the scoring script.[1]

The earlier work Kocmi and Federmann (2023a) (GEMBA-DA) adopted a straightforward methodology of assessing single score values for each segment without specifying the scale in detail. Employing a zero-shot approach, their technique showed an unparalleled accuracy in assessment, surpassing all other non-LLM metrics on the WMT22 metrics test set (Freitag et al., 2022).

Next, Lu et al. (2023) (EAPrompt) investigated prompting LLMs to assess individual error classes from a multidimensional quality metrics (MQM) framework (Freitag et al., 2021), where each error can be classified into various error classes (such as accuracy, fluency, style, terminology, etc.), subclasses (accuracy > mistranslation), and is marked with its severity (critical, major, minor). Segment scores are computed by aggregating errors, each weighted by its respective severity coefficient (25, 5, 1). While their approach employed a few-shot prompting with a chain-of-thought strategy (Wei et al., 2022), our GEMBA-MQM approach differs in two aspects: 1) We streamline the process using only single-step prompting, and 2) our prompts are universally applicable across languages, avoiding the need for manual prompt preparation for each language pair.

Another notable effort by Fernandes et al. (2023) paralleled the EAPrompt approach, also marking MQM error spans. In contrast, their approach used a PaLM-2 model, pooling MQM annotations to sample a few shot examples for the prompt. Their

[1] https://github.com/MicrosoftTranslator/GEMBA/

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 768–775
December 6–7, 2023.

```
(System) You are an annotator for the quality of machine translation. Your task is to identify
errors and assess the quality of the translation.

(user) {source_language} source:\n
```{source_segment}```\n
{target_language} translation:\n
```{target_segment}```\n
\n
Based on the source segment and machine translation surrounded with triple backticks, identify
error types in the translation and classify them. The categories of errors are: accuracy
(addition, mistranslation, omission, untranslated text), fluency (character encoding, grammar,
inconsistency, punctuation, register, spelling),
 locale convention (currency, date, name, telephone, or time format)
style (awkward), terminology (inappropriate  for context, inconsistent use), non-translation,
other, or no-error.\n
Each error is classified as one of three categories: critical, major, and minor.
Critical errors inhibit comprehension of the text. Major errors disrupt the flow, but what
the text is trying to say is still understandable. Minor errors are technically errors,
but do not disrupt the flow or hinder comprehension.

(assistant) {observed error classes}
```

Figure 1: The general prompt for GEMBA-MQM omits the gray part which performed subpar on internal data (we include it in GEMBA-locale-MQM). The "(user)" and "(assistant)" section is repeated for each few-shot example.

fine-tuning experiments did not improve system-level performance for the top-tier models.

2 Description

Our technique adopts few-shot learning with the GPT-4 model (OpenAI, 2023), prompting the model to mark quality error spans using the MQM framework. The underlying prompt template is modeled on guidelines for human annotators and shown in Figure 1.

In contrast to other methods, we use three predetermined examples (see Appendix A), allowing the method to be used with any language pair, avoiding the need to create language pair specific MQM few-shot examples. This was the original limitation that prevented Fernandes et al. (2023) from evaluating AutoMQM beyond two language pairs. Our decision was not driven by a desire to enhance performance — since domain and language-specific prompts typically boost it (Moslem et al., 2023) — but rather to ensure our method can be evaluated across any language pairs.

3 Experiments

To measure the performance of the GEMBA-MQM metric, we follow the methodology and use test data provided by the WMT22 Metrics shared task (Freitag et al., 2022) which hosts an annual evaluation of automatic metrics, benchmarking them against human gold labels.

We compare our method against the best-performing reference-based metrics of WMT22: MetrixX_XXL (non-public metric), COMET-22 (Rei et al., 2022), UNITE (Wan et al., 2022b), BLEURT-20 (Pu et al., 2021), and COMET-20 (Rei et al., 2020). In addition, we also compare against "classic" string-based metrics BLEU (Papineni et al., 2002) and ChrF (Popović, 2015). Lastly, we compare against reference-less metrics of WMT22: CometKIWI (Rei et al., 2022), Unite-src (Wan et al., 2022a), Comet-QE (Rei et al., 2021), MS-COMET-QE-22 (Kocmi et al., 2022b).

We contrast our work with other LLM-based evaluation methods such as GEMBA-DA (Kocmi and Federmann, 2023b) and EAPrompt (Lu et al., 2023), conducting experiments using two GPT models: GPT-3.5-Turbo and the more powerful GPT-4 (OpenAI, 2023).

3.1 Test set

The main evaluation of our work has been done on the MQM22 (Freitag et al., 2022) and internal Microsoft data. Furthermore, a few days before the camera-ready deadline, organizers of Metrics 2023 (Freitag et al., 2023) released results on the blind test set, showing performance on unseen data.

The MQM22 test set contains human judgments for three translation directions: English into German, English into Russian, and Chinese into English. The test set contains a total of 54 machine translation system outputs or human translations. It

contains a total of 106k segments. Translation systems are mainly from participants of the WMT22 General MT shared task (Kocmi et al., 2022a). The source segments and human reference translations for each language pair contain around 2,000 sentences from four different text domains: news, social, conversational, and e-commerce. The gold standard for scoring translation quality is based on human MQM ratings, annotated by professionals who mark individual errors in each translation, as described in Freitag et al. (2021).

The MQM23 test set is the blind set for this year's WMT Metrics shared task prepared in the same way as MQM22, but with unseen data for all participants, making it the most reliable evaluation as neither participants nor LLM could overfit to those data. The main difference from last year's iteration is the replacement of English into Russian with Hebrew into English. Also, some domains have been updated; see Kocmi et al. (2023).

Additionally, we evaluated GEMBA-MQM on a large internal test set, an extended version of the data set described by Kocmi et al. (2021). This test set contains human scores collected with source-based Direct Assessment (DA, Graham et al., 2013) and its variant DA+SQM (Kocmi et al., 2022a). This test set contains 15 high-resource languages paired with English. Specifically, these are: Arabic, Czech, Dutch, French, German, Hindi, Italian, Japanese, Korean, Polish, Portuguese, Russian, Simplified Chinese, Spanish, and Turkish.

3.2 Evaluation methods

The main use case of automatic metrics is system ranking, either when comparing a baseline to a new model, when claiming state-of-the-art results, when comparing different model architectures in ablation studies, or when deciding if to deploy a new model to production. Therefore, we focus on a method that specifically measures this target: system-level pairwise accuracy (Kocmi et al., 2021).

The pairwise accuracy is defined as the number of system pairs ranked correctly by the metric with respect to the human ranking divided by the total number of system pair comparisons.

Formally:

$$\text{Accuracy} = \frac{|\text{sign}(\text{metric}\Delta) == \text{sign}(\text{human}\Delta)|}{|\text{all system pairs}|}$$

We reproduced all scores reported in the WMT22 Metrics shared task findings paper using the official WMT22 script.[2] Reported scores match Table 11 of the WMT22 metrics findings paper (Freitag et al., 2022).

Furthermore, organizers of Metrics shared task 2023 defined a new meta-evaluation metric based on four different scenarios, each contributing to the final score with a weight of 0.25:

- system-level pairwise accuracy;
- system-level Pearson correlation;
- segment-level Accuracy-t (Deutsch et al., 2023); and
- segment-level Pearson correlation.

The motivation is to measure metrics in the most general usage scenarios (for example, for segment-level filtering) and not just for system ranking. However, we question the decision behind the use of Pearson correlation, especially on the system level. As Mathur et al. (2020) showed, Pearson used for metric evaluation is sensitive when applied to small sample sizes (in MQM23, the sample size is as little as 12 systems); it is heavily affected by outliers (Osborne and Overbay, 2004; Ma et al., 2019), which need to be removed before running the evaluation; and it measures linear correlation with the gold MQM data, which are not necessarily linear to start with (especially the discrete segment-level scores, with error weights of 0.1, 1, 5, 25).

Although it is desirable to have an automatic metric that correlates highly with human annotation behaviour and which is useful for segment-level evaluation, more research is needed regarding the proper way of testing these properties.

4 Results

In this section, we discuss the results observed on three different test sets: 1) MQM test data from WMT, 2) internal test data from Microsoft, and 3) a subset of the internal test data to measure the impact of the MQM locale convention.

4.1 Results on MQM Test Data from WMT

The results of the blind set MQM23 in Table 1 show that GEMBA-MQM outperforms all other techniques on the three languages evaluated in the system ranking scenario. Furthermore, when evaluated in the meta-evaluation scenario it achieves the third cluster rank.

In addition to the official results, we also test on MQM22 test data and show results in Table 2. The

[2] https://github.com/google-research/mt-metrics-eval

Metric	Acc.
EAPrompt-Turbo	90.9%
GEMBA-DA-GPT4	89.8%
GEMBA-locale-MQM-Turbo	89.8%
EAPrompt-Turbo	89.4%
GEMBA-MQM-GPT4	89.4%
GEMBA-DA-GPT4	87.6%
GEMBA-DA-Turbo	86.9%
GEMBA-MQM-Turbo	86.5%
GEMBA-DA-Turbo	86.5%
MetricX_XXL	85.0%
BLEURT-20	84.7%
COMET-22	83.9%
COMET-20	83.6%
UniTE	82.8%
COMETKiwi	78.8%
COMET-QE	78.1%
BERTScore	77.4%
UniTE-src	75.9%
MS-COMET-QE-22	75.5%
chrF	73.4%
BLEU	70.8%

Table 2: The system-level pairwise accuracy results for the WMT 22 metrics task test set. Gray metrics need reference translations which are not the focus of the current evaluation.

# of system pairs (N)	15 langs 4,468	Cs + De 734
COMETKiwi	79.9	81.3
GEMBA-locale-MQM-Turbo	78.6	81.3
GEMBA-MQM-Turbo	78.4	83.0
COMET-QE	77.8	79.8
COMET-22	76.5	79.2
COMET-20	76.3	79.6
BLEURT-20	75.8	79.7
chrF	68.1	70.6
BLEU	66.8	68.9

Table 3: System-level pairwise accuracy results for our internal test set. The first column is for all 15 languages, and the second is Czech and German only. All languages are paired with English.

Source	Vstupné do památky činí 16,50 Eur.
Hypothesis	Admission to the monument is 16.50 Euros.
GPT annot.	locale convention/currency: "euros"

Table 4: An example of a wrong error class "locale convention" as marked by GEMBA-locale-MQM. The translation is correct, however, we assume that the GPT model might not have liked the use of Euros in a Czech text because Euros are not used in the Czech Republic.

main conclusion is that all GEMBA-MQM variants outperform traditional metrics (such as COMET or Metric XXL). When focusing on the quality estimation task, we can see that the GEMBA-locale-MQM-Turbo method slightly outperforms EAPrompt, which is the closest similar technique.

However, we can see that our final technique GEMBA-MQM is performing significantly worse than the GEMBA-locale-MQM metric, while the only difference is the removal of the locale convention error class. We believe this to be caused by the test set. We discuss our decision to remove the locale convention error class in Section 4.3.

4.2 Results on Internal Test Data

Table 3 shows that GEMBA-MQM-Turbo outperforms almost all other metrics, losing only to COMETKIWI-22. This shows some limitations of GPT-based evaluation on blind test sets. Due to access limitations, we do not have results for GPT-4, which we assume should outperform the GPT-3.5 Turbo model. We leave this experiment for future work.

4.3 Removal of Locale Convention

When investigating the performance of GEMBA-locale-MQM on a subset of internal data (Czech and German), we observed a critical error in this prompt regarding the "locale convention" error class. GPT assigned this class for errors not related to translations. It flagged Czech sentences as a locale convention error when the currency Euro was mentioned, even when the translation was fine, see example in Table 4. We assume that it was using this error class to mark parts not standard for a given language but more investigation would be needed to draw any deeper conclusions.

The evaluation on internal test data in Table 4 showed gains of 1.7% accuracy. However, when evaluating over 15 languages, we observed a small degradation of 0.2%. For MQM22 in Table 2, the degradation is even bigger.

When we look at the distribution of the error classes over the fifteen highest resource languages in Table 5, we observe that 32% of all errors for GEMBA-locale-MQM are marked as a locale convention suggesting a misuse of GPT for this error class. Therefore, instead of explaining this class in the prompt, we removed it. This resulted in about half of the original locale errors being reassigned to other error classes, while the other half was not marked.

In conclusion, we decided to remove this class as it is not aligned with what we expected to measure and how GPT appears to be using the classes. Thus, we force GPT to classify those errors using other error categories. Given the different behaviour for internal and external test data, this deserves more

Error class	GEMBA-locale-MQM	GEMBA-MQM
accuracy	960,838 (39%)	1,072,515 (51%)
locale con.	808,702 (32%)	(0%)
fluency	674,228 (27%)	699,037 (33%)
style	23,943 (1%)	41,188 (2%)
terminology	17,379 (1%)	290,490 (14%)
Other errors	4,126 (0%)	10615 (1%)
Total	2,489,216	2,113,845

Table 5: Distribution of errors for both types of prompts over all segments of the internal test set for the Turbo model.

investigation in future work.

5 Caution with "Black Box" LLMs

Although GEMBA-MQM is the state-of-the-art technique for system ranking, we would like to discuss in this section the inherent limitations of using "black box" LLMs (such as GPT-4) when conducting academic research.

Firstly, we would like to point out that GPT-4 is a proprietary model, which leads to several problems. One of them is that we do not know which training data it was trained on, therefore any published test data should be considered as part of their training data (and is, therefore, possibly tainted). Secondly, we cannot guarantee that the model will be available in the future, or that it won't be updated in the future, meaning any results from such a model are relevant only for the specific sampling time. As Chen et al. (2023) showed, the model's performance fluctuated and decreased over the span of 2023.

As this impacts all proprietary LLMs, we advocate for increased research using publicly available models, like LLama 2 (Touvron et al., 2023). This approach ensures future findings can be compared both to "black box" LLMs while also allowing comparison to "open" models.[3]

6 Conclusion

In this paper, we have introduced and evaluated the GEMBA-MQM metric, a GPT-based metric for translation quality error marking. This technique takes advantage of the GPT-4 model with a fixed three-shot prompting strategy. Preliminary results show that GEMBA-MQM achieves a new state of the art when used as a metric for system ranking, outperforming established metrics such as COMET and BLEURT-20.

We would like to acknowledge the inherent limitations tied to using a proprietary model like GPT. Our recommendation to the academic community is to be cautious with employing GEMBA-MQM on top of GPT models. For future research, we want to explore how our approach performs with other, more open LLMs such as LLama 2 (Touvron et al., 2023). Confirming superior behaviour on publicly distributed models (at least their binaries) could open the path for broader usage of the technique in the academic environment.

Limitations

While our findings and techniques with GEMBA-MQM bring promising advancements in translation quality error marking, it is essential to highlight the limitations encountered in this study.

- Reliance on Proprietary GPT Models: GEMBA-MQM depends on the GPT-4 model, which remains proprietary in nature. We do not know what data the model was trained on or if the same model is still deployed and therefore the results are comparable. As Chen et al. (2023) showed, the model's performance fluctuated throughout 2023;

- High-Resource Languages Only: As WMT evaluations primarily focus on high-resource languages, we cannot conclude if the method will perform well on low-resource languages.

Acknowledgements

We are grateful to our anonymous reviewers for their insightful comments and patience that have helped improve the paper. We would like to thank our colleagues on the Microsoft Translator research team for their valuable feedback.

References

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. How is chatgpt's behavior changing over time? *arXiv preprint arXiv:2307.09009*.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Modifying kendall's tau for modern metric meta-evaluation. *arXiv preprint arXiv:2305.14324*.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André FT Martins, Graham Neubig,

[3]Although LLama 2 is not fully open, its binary files have been released. Thus, when used it as a scorer, we are using the exact same model.

Ankush Garg, Jonathan H Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. *arXiv preprint arXiv:2308.07286*.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Nitika Mathur, Chi kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frédéric Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of wmt23 metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and Maja Popović. 2023. Findings of the 2023 conference on machine translation (WMT23). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022a. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023a. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Tom Kocmi, Hitokazu Matsushita, and Christian Federmann. 2022b. MS-COMET: More and better human judgements improve metric performance. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 541–548, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2023. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

OpenAI. 2023. Gpt-4 technical report.

Jason W Osborne and Amy Overbay. 2004. The power of outliers (and why researchers should always check for them). *Practical Assessment, Research, and Evaluation*, 9(1):6.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Yu Wan, Keqin Bao, Dayiheng Liu, Baosong Yang, Derek F. Wong, Lidia S. Chao, Wenqiang Lei, and Jun Xie. 2022a. Alibaba-translate China's submission for WMT2022 metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 586–592, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 2022b. UniTE: Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127, Dublin, Ireland. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

A Three examples Used for Few-shot Prompting

```
English source: I do apologise about this, we must gain permission from the account holder to discuss
an order with another person, I apologise if this was done previously, however, I would not be able
to discuss this with yourself without the account holders permission.
German translation: Ich entschuldige mich dafür, wir müssen die Erlaubnis einholen, um eine Bestellung
mit einer anderen Person zu besprechen. Ich entschuldige mich, falls dies zuvor geschehen wäre, aber
ohne die Erlaubnis des Kontoinhabers wäre ich nicht in der Lage, dies mit dir involvement.
MQM annotations:
Critical:
no-error
Major:
accuracy/mistranslation - "involvement"
accuracy/omission - "the account holder"
Minor:
fluency/grammar - "wäre"
fluency/register - "dir"

English source: Talks have resumed in Vienna to try to revive the nuclear pact, with both sides
trying to gauge the prospects of success after the latest exchanges in the stop-start negotiations.
Czech transation: Ve Vídni se ve Vídni obnovily rozhovory o oživení jaderného paktu, přičemže obě
partaje se snaží posoudit vyhlídky na úspěch po posledních výměnách v jednáních.
MQM annotations:
Critical:
no-error
Major:
accuracy/addition - "ve Vídni"
accuracy/omission - "the stop-start"
Minor:
terminology/inappropriate for context - "partaje"

Chinese source: 大众点评乌鲁木齐家居商场频道为您提供高铁居然之家地址，电话，营业时间等最新商户信息，
找装修公司，就上大众点评
English translation: Urumqi Home Furnishing Store Channel provides you with the latest business
information such as the address, telephone number, business hours, etc., of high-speed rail, and
find a decoration company, and go to the reviews.
MQM annotations:
Critical:
accuracy/addition - "of high-speed rail"
Major:
accuracy/mistranslation - "go to the reviews"
Minor:
style/awkward - "etc.,"
```

Figure 2: Three examples used for all languages.

Metric Score Landscape Challenge (MSLC23): Understanding Metrics' Performance on a Wider Landscape of Translation Quality

Chi-kiu Lo 羅致翹 **Samuel Larkin** **Rebecca Knowles**
Digital Technologies Research Centre
National Research Council Canada (NRC-CNRC)
{chikiu.lo,samuel.larkin,rebecca.knowles}@nrc-cnrc.gc.ca

Abstract

The Metric Score Landscape Challenge (MSLC23) dataset aims to gain insight into metric scores on a broader/wider landscape of machine translation (MT) quality. It provides a collection of low- to medium-quality MT output on the WMT23 general task test set. Together with the high quality systems submitted to the general task, this will enable better interpretation of metric scores across a range of different levels of translation quality. With this wider range of MT quality, we also visualize and analyze metric characteristics beyond just correlation.

1 Introduction

Under time and human resource constraints, automatic metrics are often used as a proxy of manual evaluation for machine translation (MT) quality. The WMT Metrics shared task evaluates how well a variety of automatic metrics correspond to human judgments of MT quality, as evaluated on the WMT General (formerly News) shared task data. Those MT systems being evaluated are typically high-performing systems, especially for high-resource language pairs. However, in practice, the lessons learned are applied to a broader range of systems in development, including low-resource and low-quality output.

This challenge set[1] aims to gain insight into metric scores across a broader MT quality landscape. It provides a collection of low- to medium-quality MT output on the WMT23 general task test set. This serves several purposes. Together with the high quality systems submitted to the general task, this will enable more thorough understanding of metric scores across a range of different levels of translation quality: useful knowledge for researchers considering applying these metrics to lower-resource language pairs or lower-performing domains. This challenge set also allows us to explore metric characteristics beyond just correlation, which has been a main focus of past work. By expanding the range of MT quality analyzed, we shed light on some unexpected or under-explored properties of metrics, such as metrics that can distinguish between high quality systems but are not able to differentiate different levels of MT quality on the lower end of the quality scale (or vice versa) and metrics that use their space of scores in very different ways (e.g., discretized, or with specific score ranges with particularly large numbers of ties).

We focus on four language pairs: Chinese→English (ZH→EN), Hebrew↔English (HE↔EN), and English→German (EN→DE). Three of these correspond to the focus languages of the WMT 2023 Metrics shared task (EN→DE, HE→EN, ZH→EN), and they also cover several language families and aspects of translation evaluation (i.e., the paragraph-level evaluation of EN→DE), as well as including a sentence-level out-of-English direction (EN→HE). We combine source and reference data from the news portion of the WMT 2023 General MT task test sets with our challenge set, the low- and medium-quality MT output that we generated to cover a range of MT quality.

We begin by describing the training data (Section 3.1) and models (Section 3.2) used in for constructing our challenge set (Section 3.3). We also briefly describe the additional data (Section 4) and metrics (Section 5) analyzed. In Section 6 we analyze the distribution of different metrics over the challenge set. We find that some metrics exhibit strikingly different characteristics on the low-quality systems as compared to the systems submitted to WMT, while others exhibit unexpected characteristics (e.g., large numbers of tied scores) that would not have been apparent from standard correlational analysis or from high-quality WMT submitted systems alone. We conclude by arguing

[1] Available at https://github.com/nrc-cnrc/MSLC23

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 776–799
December 6–7, 2023.

that examining metric characteristics and performance over a wider landscape of MT quality—or indicating clearly when a metric has only been tested on high-quality MT—is an important factor for researchers to consider when building, presenting, and applying new metrics (especially if those metrics will be applied to lower-quality outputs).

2 Related Work

Przybocki et al. (2009) outlined four objectives in the search for new and improved automatic MT evaluation metrics: 1) "high correlation with human assessments of translation quality"; 2) "applicable to multiple target languages"; 3) "ability to differentiate between systems of varying quality" and finally, 4) "intuitive interpretation". Over the years, the WMT Metrics shared tasks (Callison-Burch et al., 2007; Bojar et al., 2017b; Freitag et al., 2021, 2022, i.a.) focused mainly on evaluating MT evaluation metrics on the first two objectives.

Many other research efforts on meta-evaluation of metrics also focused on their ability to correlate with human judgment. Graham and Baldwin (2014) introduced Williams' significance tests for understanding the confidence of the correlation analysis. Mathur et al. (2020) pointed out that Pearson's correlation is sensitive to outliers and proposed to remove outliers in Pearson's correlation analysis at system level. Kocmi et al. (2021) proposed to use pairwise accuracy to evaluate metrics based on whether the metric's pairwise rankings of two systems agrees with human pairwise rankings. Deutsch et al. (2023) introduced a tie calibration procedure enabling fair comparison between metrics that do and do not predict ties for pairwise accuracy analysis at the segment level. Marie (2022) and Lo et al. (2023) studied the relationship of metrics' score differences and statistical significance of ranking decision. Notably, these works are mostly based on the data released by WMT Metrics shared task. That means the translation output scored by the metrics in these work were generated by the participants of the WMT News/General Translation shared task, typically consisting of high-quality MT output.

There is growing interest in understanding metric performance beyond correlation. Moghe et al. (2023) note that neural metrics are not interpretable at the segment level across different language pairs. The WMT Metrics shared task introduced the challenge sets subtask (Freitag et al., 2021, 2022) to challenge metrics on particular translation errors, including negation and polarity, word/sentence addition/omission, tokenization, punctuation, numeric expression, casing number swapping, spelling, etc., in order to shed light on metric strengths and weaknesses. The challenge sets created by Macketanz et al. (2018); Avramidis et al. (2020); Avramidis and Macketanz (2022) were more linguistically motivated and covers more than 100 phenomena, including tenses, relative clauses, idioms, focus particles, etc. The ACES challenge set (Amrhein et al., 2022) covers 146 translation directions and 68 types of errors, ranging from simple perturbations to more complex errors based on discourse and real-world knowledge. The SMAUG challenge set (Alves et al., 2022) and the HWTSC challenge set (Chen et al., 2022) focused on the robustness of metrics on translation errors involving named entities, numeric/date/time entities, etc.

We note that even as MT evaluation metrics become better at correlating with human judgment on translation quality for high-quality MT systems, metric performance may be untested on low- to medium-quality MT output. Hence, we design the MSLC23 challenge set to gain insights of metric behavior on a more complete landscape.

3 Challenge Set

The challenge set consists of data translated by MT systems of varying quality. We describe the training data used to build these systems as well as the MT models.

3.1 Training Data

To build the lower-quality MT systems that we analyze in this work, we use standard WMT datasets from WMT 2023 (Kocmi et al., 2023) for EN→DE and HE↔EN and from WMT 2017 (Bojar et al., 2017a) for ZH→EN. For EN→DE and ZH→EN, we used the *newstest2020* data as our validation set. For HE↔EN, we used a random sample of 2000 lines, ensuring no overlap between sentence pairs in the training and validation set. For full details of training data, see Appendix A. Appendix B describes the preprocessing and subword segmentation performed.

3.2 MT Models

We build two main types of systems: baselines and pseudo-low-resource systems. All systems were

built using Sockeye-3.1.31 (Hieber et al., 2022), commit 13c63be5, with PyTorch-1.12.1 (Paszke et al., 2019). For more details on parameters and training, see Appendix C.

The baselines are standard Transformer models trained over the available data, but without any additional components (e.g., backtranslation, factors, tagging, etc.). The pseudo-low-resource systems are produced using subsets of the training data, to simulate lower-resource settings (see Appendix D for details). We checkpoint all systems frequently so that we can use output at various levels of training as representative of different levels of quality.

We note that the EN→DE 2023 shared task is performed at the paragraph level. In our work we do not perform paragraph-level MT; instead we use as a baseline sentence segmentation, translation of the individual sentences, and concatenation back into paragraphs of the resulting translation output.

3.3 Translation Output

We use the news data portions of each of the 2023 General MT task test sets for these language pairs. This consists of 139 paragraphs for EN→DE (translated by 12 different systems), 516 lines for EN→HE (translated by 6 different systems), 1558 lines for HE→EN (translated by 6 different systems), and 763 lines for ZH→EN (translated by 6 different systems).

We use checkpoints from each of the systems we built to produce the low- and medium-quality MT output. For ZH→EN and HE↔EN, all checkpoints were selected from the baseline systems; for EN→DE, they were selected from the baseline system as well as the 50k, 200k, and 400k pseudo-low-resource systems. These checkpoints were selected to cover a range of BLEU scores from less than 1 to between 20 and 30 (shown in Appendix E, Tables 11 and 12; we assign the selected systems the letters A through F, or through L in the case of EN→DE, with A being the lowest quality system in all cases),[2] and were then spot-checked manually to confirm that they did generally appear to represent incremental (but noticeable) improvements in quality. We note that we did not perform a full or extensive manual evaluation, and as such cannot make claims of statistically significant human judgment differences between the checkpoints. Another potential limitation of this choice to select checkpoints is that they may be more similar to one another than separately-trained systems would be (cf. the benefits of ensembling diverse sets of systems or the potential minor drawbacks of ensembling checkpoints rather than separately trained models in Farajian et al. (2016); Sennrich et al. (2016), i.a.). Nevertheless, we expect this should provide some coverage of low- to mid-quality MT for scoring by various metrics.

[2]Computed with sacreBLEU (Post, 2018) with signature: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1 For HE↔EN, the systems were selected based on BLEU scores computed with refA as the reference, as refB had not yet been released. The range of BLEU scores remains similar, and we present all other HE↔EN results based on scores with refB.

4 General MT Submissions

Our challenge set has aimed to cover the range of low-quality MT systems, but to obtain a fuller picture, we also include the metric scores assigned to the systems submitted to the WMT2023 General MT Task (Kocmi et al., 2023) in our analysis. For these systems, we have human annotation scores in the form of multidimensional quality metrics (MQM; Burchardt, 2013) scores for EN→DE, HE→EN, and ZH→EN.

5 Metrics

There are dozens of metrics submitted by the task organizers and participants in WMT23 Metrics shared task. Under the time and space limitations, we only examine the baseline metrics submitted by the task organizers and the primary metrics submitted by the participants. Due to the random shuffling of items in the challenge sets before their delivery to the scorers, we can only examine metrics that produce scores at the segment level, as the system-level scores returned do not correspond to the underlying systems in our datasets. We describe the metrics included in this work in Appendix F.

6 Analysis

We are interested in metric performance and characteristics at both the segment level and the system level. In the case of EN→DE, the segments are paragraphs, while in all other cases they are typically sentences. For metrics that use the reference, HE↔EN are scored against refB (a higher-quality reference translation than refA), while EN→DE and ZH→EN are scored against refA.

For EN→DE, HE→EN, and ZH→EN, we have access to human scores for all submitted WMT MT systems (but not for the challenge set systems).

These take the form of MQM scores over a consistent subset of the test set. For the remainder of this work unless otherwise noted, in order to make appropriate comparisons between metric scores on the challenge set, metric scores on the submitted WMT MT systems, and the human annotations, we restrict our analysis to only those segments that are in the news domain and that correspond to the set for which we have human annotations (104 paragraphs for EN→DE, 619 segments we use all 516 segments of the test set data that are in the news domain because we do not have any human annotations).

6.1 Segment Level

Since we only have human annotations for the WMT MT submissions and not our challenge set, we must be cautious in the conclusions that we draw about metric *performance* from the scores they assign to segments. However, we can observe that different metrics exhibit different *characteristics*, even as they score an identical set of segments over an identical set of systems.

6.1.1 Distributions of Scores

As we see in the histograms along the diagonal of Figure 1, showing a subset of the baseline and submitted metrics, different metrics exhibit very different score distributions. This can also be seen in Figures 3, 4, 5, and 6 in Appendix G. Some show a somewhat bimodal distribution of scores, some are closer to normally distributed, and there are a number of metrics whose score distributions do not fall into either of those patterns. Additionally, they differ in whether they exhibit a strong separation between the segments produced by the low-quality systems from our challenge set and the segments produced by the WMT submissions or whether they assign a range of low to high scores to most systems (i.e., having clear overlap in score range across all systems). While we cannot conclude that any of these metrics is more *accurate*, we can note that their varied characteristics suggest that they may be measuring different things and/or that different metrics may have different strengths and weaknesses across the translation quality landscape.

There are also metrics that use an approximation of a discrete score space, such as GEMBA-MQM. This particular metric also scores nearly all segments produced by our low-quality systems as the lowest available score, particularly for EN→DE, meaning it would not be a suitable metric to distinguish between low-to-mid quality (e.g., low-resource) translation systems. XCOMET-Ensemble assigns a wider range of scores to the low-quality segments, but the range and distribution of those scores is fairly consistent across the low-quality systems in our challenge set, meaning that it also struggles to distinguish between system quality levels at the lower end, albeit for a different reason. We can also see this when we examine system-level scores.

6.1.2 Universal Translations and Universal Scores

In Yan et al. (2023), the authors observe what they term "universal translations": target language output that receives high scores regardless of the reference to which they are compared. Here, we observe what one might consider to be "universal scores" instead. Some metrics, like Calibri-COMET22, use a wide range of scores in general, but have a very small subset of scores that appear a very large number of times. For the set of annotated news segments across all challenge set and WMT MT systems for EN→DE, 1673 unique scores are assigned to segments. The vast majority occur only once, but there are two non-minimum/maximum scores that occur 210 and 206 times, respectively (the score zero, i.e. the maximum score for perfect translation in this case, also appears 206 times). In contrast, COMET assigns 2446 unique scores over the same subset of segments, with the most frequent of those scores occurring 7 times. We note that Calibri-COMET22 (and Calibri-COMET22-QE) exhibit this frequently-appearing-score characteristic across the different language pairs, though the number and exact value of the extremely frequent scores differs across language pairs. Importantly, this is not explainable by the data itself: other metrics assign a wide variety of scores to the same segments that receive these particularly common scores, which makes the common scores visible in the Calibri-COMET22 column as the apparent vertical lines (most visible in comparison to COMET). As is evident from both the histogram and the scatterplots, these common scores are most frequently assigned to the segments in our challenge set, to the extent that this unexpected characteristic is not clearly visible when the plot is restricted to only the WMT MT submissions rather than including the challenge set (see Figure 10). This highlights the importance of performing evaluation over a wide

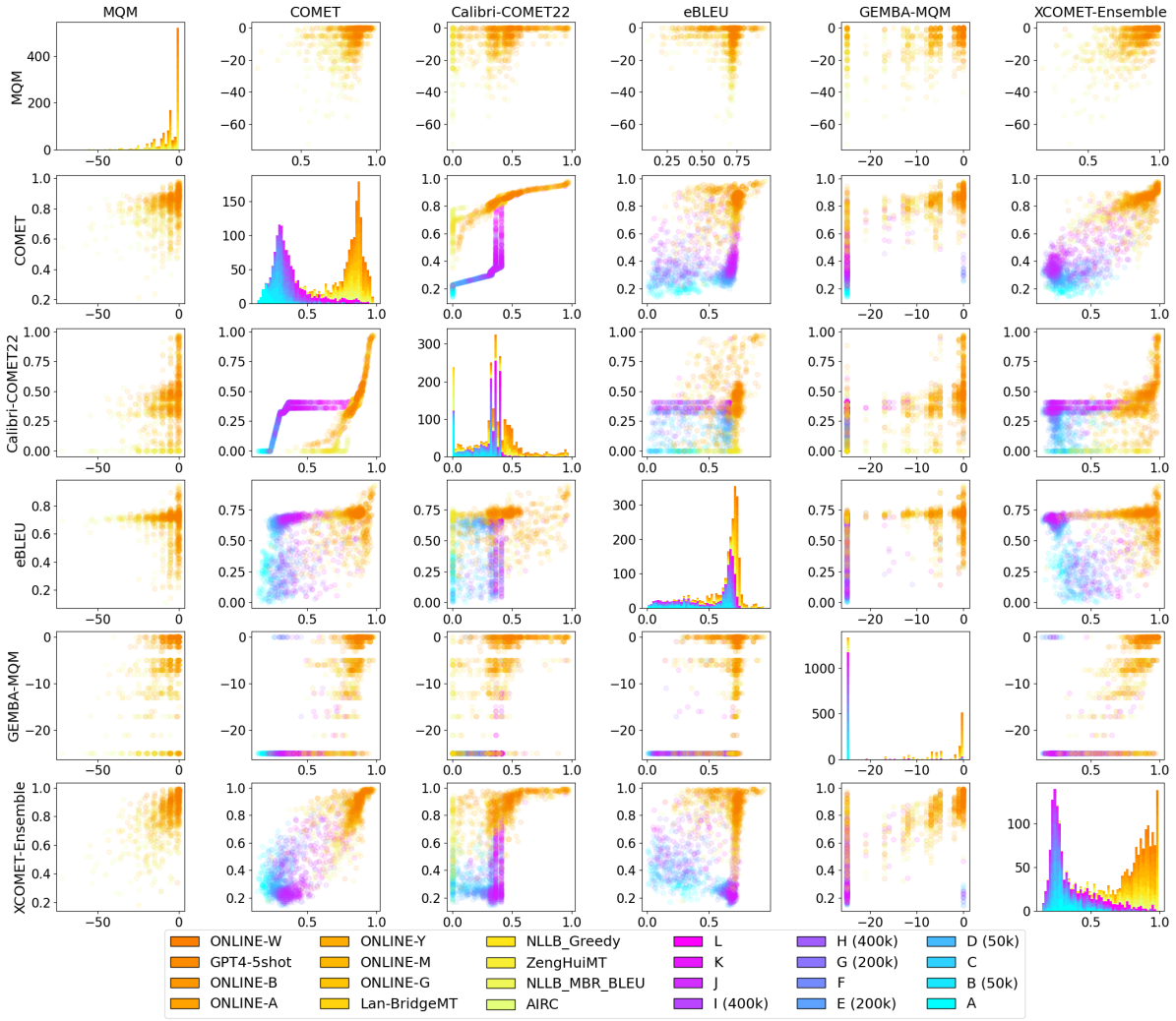


Figure 1: A subset of the metrics (and MQM scores) for EN→DE. The diagonal entries show stacked histograms of segment scores across the challenge set (cool colours/bottom) and submitted WMT systems (warm colours/top). The off-diagonal entries are scatterplots where each point is a single segment positioned according to the score assigned to it by row and column metrics; each point is coloured according to the MT system that produced it.

range of MT quality, in order to discover unexpected issues like this prior to applying the metrics to low-resource or otherwise low-quality MT.

The eBLEU metric exhibits a slightly more dispersed version of this, where a large number of segments receive scores in a fairly narrow band relative to the metric's overall score distributions. However, in the case of eBLEU, this is not specific to the challenge set data, but is also observed in the WMT MT data.

6.2 System Level

To analyze system-level scores, we produce them as an average over all of the segment-level scores in the restricted test set (news domain segments in the set of segments for which WMT MT systems were human-annotated) for a given MT system.[3] These system-level scores can also be used in order to gain a better understanding of the overall range of a metric's scores, as well as what kind of scores are assigned to very low quality machine translation (e.g., the A and B systems from the challenge set).

In Figure 2 (as well as Figures 7, 8, and 9 in Appendix G), we observe that metrics show different patterns of scores at the system level. We observe that some metrics exhibit unexpected characteristics on the low-quality data, such as MaTESe, which ranks some of the low-quality systems in reverse order.[4]

We do not have MQM scores for any of the data in the challenge set, which means that we do not know how much of a gap in quality there is between our best low-quality system and the lowest-performing MT systems submitted to WMT. However, we can observe that metrics differ widely in their estimates of the gap; embed_llama, for example, shows error bar overlap between the highest performing system from our challenge set and the lowest-scored system from the submissions, while GEMBA-MQM shows a very large gap between the two groups of systems,[5] with many of the other metrics falling between these extremes.

Similarly, again examining characteristics without making claims about metric performance, we notice variety amongst the metrics in terms of the range of scores they assigned to each group of systems, as seen in the slope of the system scores. In some cases, there are quite similar slopes (e.g., embed_llama), while in other cases there is a steep slope for the challenge set as compared to the WMT MT systems (e.g., BERTScore or COMET) meaning that the challenge set covers a wide range of (lower) scores while the WMT MT set covers a smaller range of higher scores, and finally some systems where the slopes are similar but both less steep (e.g., Calibri-COMET22-QE and GEMBA-MQM) and each set of systems covers a small range of scores with a gap in between. Without MQM scores for the challenge set, we do not know whether one of these patterns is indicative of a metric that more closely resembles human annotations or not (i.e., we do not know whether the challenge set covers a wider range of quality than the WMT MT systems, which would support metrics having a steeper slope/wider range in the scores assigned to it).

We also note some variety across language pairs. The reversal of scores seen in MaTESe is less obvious in the EN→DE data, though that may be related to greater overlap in the EN→DE challenge set data quality.

In future work, obtaining MQM scores for one or more of the systems in our challenge set would permit us to draw conclusions about metric performance in these areas (i.e., about whether there is indeed quality overlap between the two sets, and what appropriate ranges of scores might be for each of the sets).

All of these observations about variation in metric characteristics raise an important issue in the evaluation and adoption of new metrics: since their correlation with human rankings is often demonstrated on the high-quality MT output being scored at WMT, it is not necessarily appropriate to use them for the evaluation of low-resource or lower-quality MT output without additional study.

6.3 Additional Discussion

We briefly mention two other items of note from our exploration of the data.

Outside of the set of data for which there are human annotations (i.e., not appearing in our figures), for HE→EN there are 14 news domain segments for which the ZengHuiMT system output an empty string. Different metrics handle this in

[3] This includes the baseline BLEU.

[4] Though we do not have extensive human evaluation, we are confident that, e.g., system E should not be ranked below system A.

[5] For EN→DE, in Figure 7, the Calibri-COMET22 metrics both find several of the highest performing systems from our challenge set to be better than several of the submitted systems, while most other metrics rank the challenge set systems below the submitted systems.

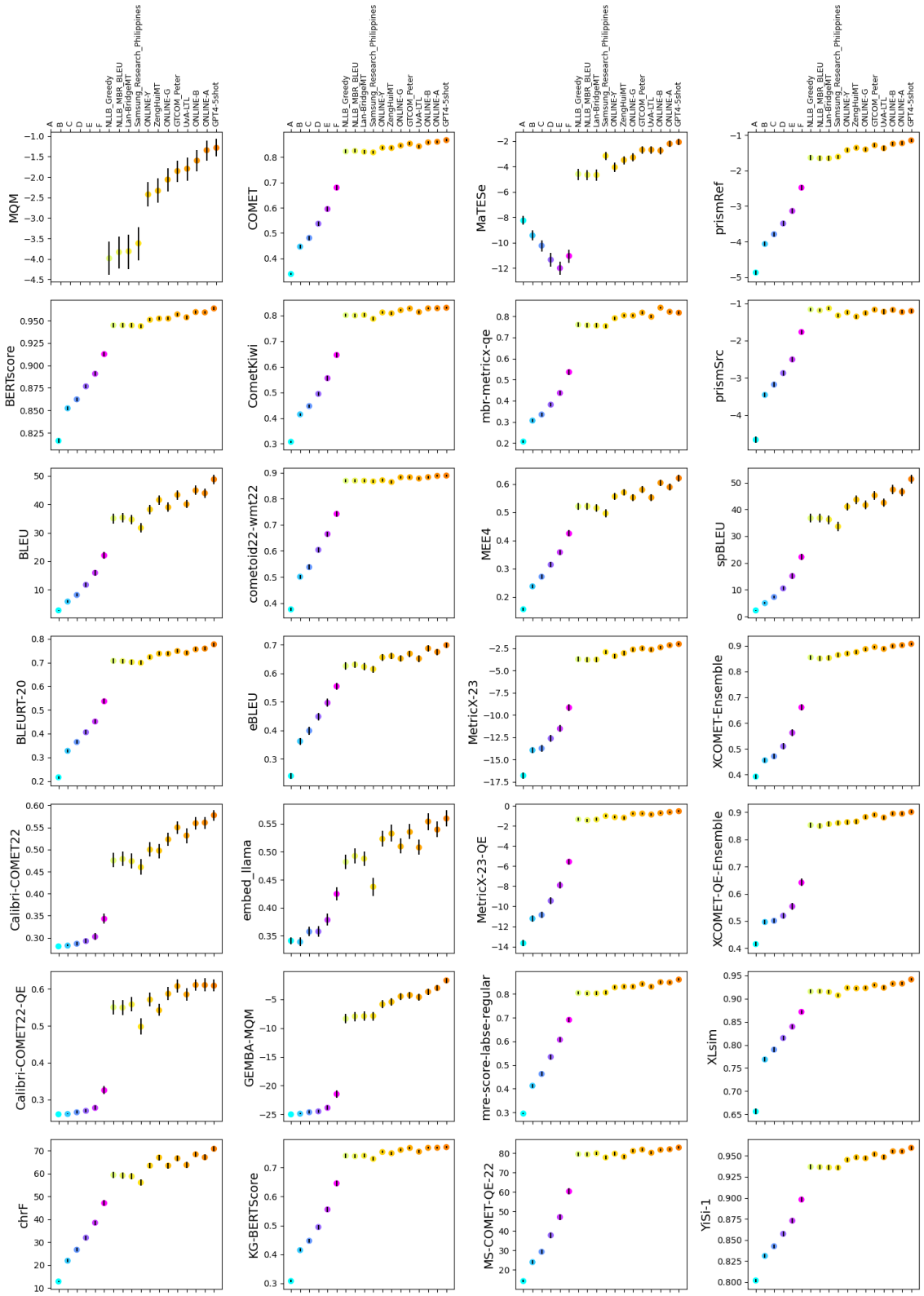


Figure 2: System average scores (with error bars computed via bootstrap resampling 1000 times for $p < 0.05$) for HE→EN across the challenge set (cool colours/left) and submitted WMT systems (warm colours/right). Our challenge set systems are ordered from left to right with BLEU scores, while the submitted WMT systems are ordered by MQM score on the news domain.

Metric	Score	Range
BERTscore	0.000	(0.000, 1.000)
BLEU	0.000	(0.000, 100.000)
BLEURT-20	0.055	(0.000, 1.030)
Calibri-COMET22	0.328	(0.000, 0.990)
Calibri-COMET22-QE	0.083	(0.000, 1.000)
chrF	0.000	(0.000, 100.000)
COMET*	0.796	(0.287, 0.995)
CometKiwi*	0.647	(0.261, 0.902)
cometoid22-wmt22	0.597	(0.268, 0.994)
eBLEU	0.000	(0.000, 1.000)
embed_llama	0.510	(0.040, 1.000)
GEMBA-MQM	-25.000	(-25.000, 0.000)
KG-BERTScore*	0.682	(0.285, 0.886)
MaTESe	0.000	(-25.000, 0.000)
mbr-metricx-qe	0.027	(-0.004, 0.998)
MEE4	0.000	(0.000, 1.000)
MetricX-23*	-25.597	(-25.618, 0.198)
MetricX-23-QE*	-24.546	(-24.557, 0.848)
mre-score-labse-regular*	0.772	(0.266, 0.965)
MS-COMET-QE-22*	59.243	(1.641, 94.075)
prismRef	-5.256	(-8.685, -0.077)
prismSrc	-6.829	(-10.027, -0.111)
spBLEU	0.000	(0.000, 100.000)
XCOMET-Ensemble*	0.917	(0.291, 0.994)
XCOMET-QE-Ensemble*	0.899	(0.290, 0.998)
XLsim	0.911	(0.569, 1.000)
YiSi-1	0.000	(0.000, 1.000)

Table 1: Average metric scores assigned to empty strings in the HE→EN news data, shown with the full range of metric scores assigned to the news data. Metrics with asterisks by their name did not assign the same scores to all the empty strings, though the differences were quite small.

different ways; some assign a score of 0 (or the metric's lower bound score), while others assigned relatively high scores due to the fact that the source and reference were very short (each source and reference consisted only of a single period). Table 1 shows the scores assigned by metrics to these empty strings as well as the range of scores over the full HE→EN news data (including segments that were not human-annotated, as the empty strings were also not included for human annotation).

We also observe two examples of systems that receive noticeably lower scores from a number of metrics than would be expected based on their human ranking: NLLB_Greedy (EN→DE, Figure 3) and Samsung_Research_Philippines (HE→EN, Figure 5). We leave this as an area for future investigation.

7 Conclusions

This challenge set expands the range of system quality scored by metrics at the shared task. This expanded range of MT quality reveals interesting characteristics and limitations of some new metrics when applied to a broader range of systems. The smaller variations in segment-level scores given by some metrics at the low end of quality could indicate that these metrics struggle to discriminate low-quality MT systems. This is further shown by the observation that some metrics rank the low-quality systems in reverse order at the system level. We have discovered a "universal score" phenomenon for some metrics, where a small subset of non-minimum/maximum distinct scores are assigned to a variety of translation output. This characteristic was not visible in the high-quality MT output alone, highlighting the importance of this type of testing. We also observe diverse behaviors from different metrics on empty string translation.

Our challenge set serves as a complement to the standard correlation-based analyses and also provides useful information to researchers who are considering using these metrics in low-resource or low-quality domains. We recommend that metric researchers check their metrics' performance on a wider landscape of translation quality or be clear about the limitations of their metrics' testing.

Limitations

A major limitation of this work is our choice to select low-quality systems on the basis of BLEU scores, which was done for reasons of time and cost. We attempted to mitigate this by spot-checking to confirm that we saw noticeable differences between various pairs of low-quality systems, but a more thorough human annotation would be beneficial. We are also limited in the set of languages we have explored, using only four language pairs, as well as the limited news domain.

Acknowledgements

We thank Adam Poliak for comments and feedback on a sample of the English–Hebrew dataset. We thank our colleagues and the anonymous reviewers for feedback and suggestions as well as the WMT Metrics Task and General Task organizers for providing data and annotations.

References

Duarte Alves, Ricardo Rei, Ana C Farinha, José G. C. de Souza, and André F. T. Martins. 2022. Robust MT evaluation with sentence-level multilingual augmentation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 469–478, Abu

Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2022. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Eleftherios Avramidis and Vivien Macketanz. 2022. Linguistically motivated evaluation of machine translation metrics based on a challenge set. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 514–529, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, Aljoscha Burchardt, and Sebastian Möller. 2020. Fine-grained linguistic evaluation for state-of-the-art machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 346–356, Online. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017a. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017b. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Aljoscha Burchardt. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Xiaoyu Chen, Daimeng Wei, Hengchao Shang, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Ting Zhu, Mengli Zhu, Ning Xie, Lizhi Lei, Shimin Tao, Hao Yang, and Ying Qin. 2022. Exploring robustness of machine translation metrics: A study of twenty-two automatic metrics in the WMT22 metric task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 530–540, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Modifying kendall's tau for modern metric meta-evaluation.

Sören DREANO, Derek Molloy, and Noel Murphy. 2023. Embed_Llama: using LLM embeddings for the Metrics Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Muhammad ElNokrashy and Tom Kocmi. 2023. eBLEU: Using Simple Word Embeddings For Efficient Machine Translation Evaluation. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

M. Amin Farajian, Rajen Chatterjee, Costanza Conforti, Shahab Jalalvand, Vevake Balaraman, Mattia A. Di Gangi, Duygu Ataman, Marco Turchi, Matteo Negri, and Marcello Federico. 2016. FBK's neural machine translation systems for IWSLT 2016. In *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C. International Workshop on Spoken Language Translation.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Thamme Gowda, Tom Kocmi, and Marcin Junczys-Dowmunt. 2023. Cometoid: Distilling Strong Reference-based Machine Translation Metrics into Even Stronger Quality Estimation Metrics. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Yvette Graham and Timothy Baldwin. 2014. Testing for significance of increased correlation with human judgment. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 172–176, Doha, Qatar. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Meghan Dowling, Maria Eskevich, Teresa Lynn, and Lamia Tounsi. 2016. Is all that glitters in machine translation quality

estimation really gold? In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3124–3134, Osaka, Japan. The COLING 2016 Organizing Committee.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2014. Is machine translation getting better over time? In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 443–451, Gothenburg, Sweden. Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023. xcomet: Transparent machine translation evaluation through fine-grained error detection.

Felix Hieber, Michael Denkowski, Tobias Domhan, Barbara Darques Barros, Celina Dong Ye, Xing Niu, Cuong Hoang, Ke Tran, Benjamin Hsu, Maria Nadejde, Surafel Lakew, Prashant Mathur, Anna Currey, and Marcello Federico. 2022. Sockeye 3: Fast neural machine translation with pytorch.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google Submission to the WMT 2023 Metrics Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Christof Monz, Makoto Morishita, Murray Kenton, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23). In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting Quality Error Spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Tom Kocmi, Hitokazu Matsushita, and Christian Federmann. 2022. MS-COMET: More and better human judgements improve metric performance. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 541–548, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Chi-kiu Lo, Rebecca Knowles, and Cyril Goutte. 2023. Beyond correlation: Making sense of the score differences of new mt evaluation metrics. In *Proceedings of Machine Translation Summit XIX Vol. 1: Research Track*, pages 186–199.

Vivien Macketanz, Renlong Ai, Aljoscha Burchardt, and Hans Uszkoreit. 2018. TQ-AutoTest – an automated test suite for (machine) translation quality. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Benjamin Marie. 2022. Yes, we need statistical significance testing. https://pub.towardsai.net/yes-we-need-statistical-significance-testing-927a8d21f9f0.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Nikita Moghe, Tom Sherborne, Mark Steedman, and Alexandra Birch. 2023. Extrinsic evaluation of machine translation metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13060–13078, Toronto, Canada. Association for Computational Linguistics.

Anthony Moi and Nicolas Patry. 2022. Huggingface's tokenizers.

Ananya Mukherjee and Manish Shrivastava. 2023. MEE4 and XLsim : IIIT HYD's Submissions' for WMT23 Metrics Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Subhajit Naskar, Daniel Deutsch, and Markus Freitag. 2023. Quality Estimation using Minimum Bayes Risk. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Niccolò Campolungo, and Roberto Navigli. 2022. MaTESe: Machine translation evaluation as a sequence tagging problem. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 569–577, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Mark Przybocki, Kay Peterson, Sébastien Bronsart, and Gregory Sanders. 2009. The nist 2008 metrics for machine translation challenge—overview, methodology, metrics, and results. *Machine Translation*, 23(2/3):71–103.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Vasiliy Viskov, George Kokush, Daniil Larionov, Steffen Eger, and Alexander Panchenko. 2023. Semantically-Informed Regressive Encoder Score. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Zhanglin Wu, Yilun Liu, Min Zhang, Xiaofeng Zhao, Junhao Zhu, Ming Zhu, Xiaosong Qiao, Jingfei Zhang, Ma Miaomiao, Zhao Yanqing, Song Peng, Shimin Tao, Hao Yang, and Yanfei Jiang. 2023. Empowering a Metric with LLM-assisted Named Entity Annotation: HW-TSC's Submission to the WMT23 Metrics Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Yiming Yan, Tao Wang, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Mingxuan Wang. 2023. BLEURT has universal translations: An analysis of automatic metrics by minimum risk training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5428–5443, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Corpora Sizes

Corpora are from WMT23 (Kocmi et al., 2023) and WMT17 (Bojar et al., 2017a).

A.1 EN→DE

From the download table at EMNLP 2023: General Machine Translation, we retrieved all **EN→DE** corpora. The *train* corpus is composed of *airbaltic*, *czechtourism*, *ecb2017*, *EESC2017*, *EMA2016*, *rapid2016*, *europarl-v10*, *news-commentary-v18*, *WikiMatrix.v1.de-en.langid* and *wikititles-v3*. We chose *newstest2020* as our *validation*. Corpora statistics are described in Table 2.

Name	# lines	# de words	# en words
train	14,227,278	234,635,104	246,351,534
validation	1418	45,855	44,018

Table 2: Corpora sizes for EN→DE, computed on raw text (not tokenized) using `wc`.

A.2 HE↔EN

From the instructions of EMNLP 2023: General Machine Translation, we retrieved all **HE→EN** corpora.[6] Then, using *wmt23-heen/train.{heb,eng}* and *WikiMatrix.en-he*, we sampled sentence pairs for the *validation* and *test* sets and the remaining pairs were used for *train* making sure that all three are mutually exclusive. Corpora statistics are described in Table 3.

Name	# lines	# en words	# he words
train	2,227,830	38,307,579	30,943,929
validation	2000	20,459	16,620

Table 3: Corpora sizes for EN→HE, computed on raw text (not tokenized) using `wc`.

A.3 ZH→EN

We used corpora from WMT2017.[7] *train* is composed of all 20 *Books*, *casia2015*, *casict2015*, *casict-A*, *casict-B*, *datum*, *NEU*, *news-commentary-v18.en-zh*, *WikiMatrix.v1.en-zh.langid* and *wikititles-v3*. We chose *newstest2020* as our *validation*. Corpora statistics are described in Table 4.

Name	# lines	# en words	# zh words
train	12,995,613	218,659,998	43,676,661
validation	2000	65,561	3716

Table 4: Corpora sizes for ZH→EN, computed on raw text (not tokenized) using `wc`.

B Subword Segmentation

After some light normalization consisting of converting non-breaking hyphen, normalizing spaces, replacing control characters with spaces and collapsing multiple spaces,[8] we trained a 32k tokens, bilingual sentencepiece unigram subtokenizer using HuggingFace's tokenizers (Moi and Patry, 2022) for each language pair. The corpora used for training the subword model were:

- **EN→HE** uses all of *wmt23-heen/train.{eng,eng}*
- **EN→DE** uses our concatenated *train*
- **ZH→EN** uses our concatenated *train*

C System Descriptions

For all systems, we used Sockeye-3.1.31 (Hieber et al., 2022), commit `13c63be5` with PyTorch-1.12.1 (Paszke et al., 2019). Training was performed on 4 Tesla V100-SXM2-32GB GPUs for **EN→DE** and **ZH→EN** and 4 Tesla V100-SXM2-16GB GPUs for **EN→HE** and **HE→EN**. Training times are shown in Table 5.

Name	Time (h)
ende	6 - 35
enhe	10.3
heen	6.5
zhen	83.5

Table 5: Training times in hours.

Table 6 describes the differences with Sockeye's default parameters. Note that we kept all intermediate checkpoints (from which we later select the outputs used for the challenge set) and used the entire validation during checkpoint evaluation.

[6] `mtdata get-recipe -ri wmt23-heen -o wmt23-heen`

[7] https://www.statmt.org/wmt17/translation-task.html

[8] https://github.com/nrc-cnrc/PortageTextProcessing/blob/main/bin/clean-utf8-text.pl

Name	Value
amp	*True*
grading clipping type	*abs*
max sequence length	*200:200*
batch type	*max-word*
checkpoint interval	*10*
initial learning rate	*0.06325*
learning rate scheduler type	*inv-sqrt-decay*
learning rate warmup	*4000*
max checkpoints	*110*
max epochs	*1000*
max num checkpoint not improved	*32*
optimizer	*Adam*
optimizer Betas	*0.9, 0.98*
optimized metric	*BLEU*
update interval	*10*
attention heads	*16:16*
shared vocabulary	*True*
transformer FFN	*4096:4096*
transformer model size	*1024:1024*
weight tying	*True*

Table 6: Differences from Sockeye's default parameters.

On top of the changes from Table 6, for **EN→HE** and **HE→EN**, we lowered the **batch size** to *6144* and changed **max checkpoints** to *330*.

For all language pairs, we have trained a baseline system using the entire *train* corpus. Additionally, for **EN→DE**, we also trained systems that use a uniformly random subsample of *train* namely, 50k, 200k and 400k (the pseudo-low-resource systems).

D Pseudo-Low-Resource Corpora

Due to human error in the sampling code, the pseudo-low-resource training data used for the EN→DE systems trained on 50k, 200k, and 400k—intended to be a random sample from the full training data—instead primarily consists of data from the first four corpora shown in Table 7. Table 8 shows the small number of differences between these subsampled corpora and simply selecting the first n lines of the full training corpus. The main consequence of this is that these systems may be skewed towards particular domains.

E Checkpoints in Challenge Set

In Table 9 we see the checkpoint IDs for systems included in the challenge set for HE↔EN and ZH→EN. Table 10 shows the same for EN→DE. The corresponding BLEU scores are shown in Tables 11 and 12, respectively.

Corpus Name	# Sentences
airbaltic	839
czechtourism	6758
ecb2017	4147
EESC2017	2,857,850
EMA2016	347,631
rapid2016	1,030,808
europarl-v10	828,473
news-commentary-v18	203,744
WikiMatrix.v1	2,579,106
wikititles	1,474,203
total	14,227,278

Table 7: (EN→DE) Sub-corpora sizes in the order they were merged to create the final sampled *train*.

Sample Size	# Differences	# lines *EESC2017*
50k	282	38,256
200k	70	188,256
400k	34	388,256

Table 8: EN→DE; Number of sentences that are different from the original train's head and how many sentences from *EESC2017* that were used.

System	EN→HE	HE→EN	ZH→EN
A	68	58	61
B	98	87	91
C	115	102	115
D	135	117	139
E	171	140	222
F	392	219	480

Table 9: Checkpoint IDs for systems included in challenge set (HE↔EN and ZH→EN).

System	EN→DE
A	54
B (50k)	1
C	79
D (50k)	7
E (200k)	2
F	91
G (200k)	27
H (400k)	4
I (400k)	43
J	102
K	129
L	313

Table 10: Checkpoint IDs for systems included in challenge set (EN→DE); parenthetical numbers indicate one of the pseudo-low-resource systems rather than the full training data system.

System	EN→HE	HE→EN	ZH→EN
A	0.6	0.7	0.9
B	3.1	4.3	5.0
C	7.2	7.3	9.3
D	11.4	11.4	13.1
E	16.6	16.0	18.5
F	26.2	23.9	23.2

Table 11: BLEU scores for systems included in challenge set over the full news data in the challenge set (HE↔EN computed with refB).

System	EN→DE
A	0.7
B (50k)	2.5
C	4.2
D (50k)	4.7
E (200k)	4.7
F	8.9
G (200k)	9.5
H (400k)	10.4
I (400k)	12.0
J	12.8
K	18.7
L	29.9

Table 12: BLEU scores for systems included in challenge set (EN→DE) over the full news data in the challenge set; parenthetical numbers indicate one of the pseudo-low-resource systems rather than the full training data system.

F Metrics

Table 13 shows a summary of the human annotations and metrics included in this work and the translation directions they participated in. In the following, we briefly describe the key characteristic of each metric.

F.1 Baseline Metrics

BLEU (Papineni et al., 2002) is the (clipped) precision of word n-grams between the MT output and its reference weighted by a brevity penalty.

spBLEU (Team et al., 2022) is BLEU computed with subword tokenization done by the Flores-200 Sentencepiece Model (Kudo and Richardson, 2018).

chrF (Popović, 2015) uses character n-grams to compare the MT output with the reference and it is a balance of precision and recall.

BERTScore (Zhang et al., 2020) uses cosine similarity of contextual embeddings from pre-trained transformers to compute F-scores of sentence level similarity.

BLEURT-20 (Sellam et al., 2020) is fine-tuning RemBERT to predict direct assessment (DA; Graham et al., 2013, 2014, 2016) scores for a MT-reference pair.

COMET (COMET-22) (Rei et al., 2022) is an ensemble of two models: COMET-20 and a multitask model jointly predicting sentence-level multidimensional quality metrics (MQM) and word-level translation quality annotation, where COMET-20 is fine-tuning XLM-R to predict DA scores for a MT-source-reference tuple. **CometKiwi** is a quality estimation metric that is similar to COMET, except it scores the MT output against the source, instead of the reference translation.

MS-COMET-QE-22 (Kocmi et al., 2022) is a COMET-QE-20 based quality estimation metric trained on a larger and filtered set of human judgements, covering 113 languages and 15 domains.

prismRef (Thompson and Post, 2020) uses a neural paraphrase model to score the MT output against the reference translation. **prismSrc** is the quality estimation version, which scores the MT output against the source, instead of the reference translation.

YiSi-1 (Lo, 2019) measures the semantic similarity between the MT output and reference by the IDF-weighted cosine similarity of contextual embeddings extracted from pretrained language models, e.g. RoBERTa, CamemBERT, XLM-R, etc., depending on the target language in evaluation.

F.2 Primary submissions

Calibri-COMET22 uses isotonic regression on the COMET-22 output scores to predict the fraction of translations with no error produced by the MT system. **Calibri-COMET22-QE** is a quality estimation metric that is similar to Calibri-COMET22, where it uses COMETKiwi as base model.

cometoid22-wmt22 (Gowda et al., 2023) is a quality estimation metric that uses COMET-22 as a teacher metric and trains a student model to predict the teacher scores without using reference translation.

eBLEU (ElNokrashy and Kocmi, 2023) uses non-contextual word embeddings and relative meaning diffusion tensors to approximate the token similarity in the MT output and reference and computes translation quality scores similar to BLEU.

embed_llama (DREANO et al., 2023) is the

Metric Name	EN→DE	EN→HE	HE→EN	ZH→EN	Reference-based
Human annotation					
MQM	✓		✓	✓	
Metrics					
BERTScore	✓	✓	✓	✓	✓
BLEU	✓	✓	✓	✓	✓
BLEURT-20	✓	✓	✓	✓	✓
Calibri-COMET22	✓	✓	✓	✓	✓
Calibri-COMET22-QE	✓	✓	✓	✓	
chrF	✓	✓	✓	✓	✓
COMET	✓	✓	✓	✓	✓
CometKiwi	✓	✓	✓	✓	
cometoid22-wmt22	✓	✓	✓	✓	
eBLEU	✓	✓	✓	✓	✓
embed_llama	✓	✓	✓	✓	✓
GEMBA-MQM	✓	✓	✓	✓	
KG-BERTScore	✓	✓	✓	✓	
MaTESe	✓		✓	✓	✓
mbr-metricx-qe	✓		✓	✓	
MEE4	✓	✓	✓	✓	✓
MetricX-23	✓	✓	✓	✓	✓
MetricX-23-QE	✓	✓	✓	✓	
mre-score-labse-regular	✓	✓	✓	✓	✓
MS-COMET-QE-22	✓	✓	✓	✓	
prismRef	✓	✓	✓	✓	✓
prismSrc	✓	✓	✓	✓	
spBLEU (flores-200)	✓	✓	✓	✓	✓
XCOMET-Ensemble	✓	✓	✓	✓	✓
XCOMET-QE-Ensemble	✓	✓	✓	✓	
XLsim	✓	✓	✓	✓	✓
YiSi-1	✓	✓	✓	✓	✓

Table 13: Human annotation and metrics included in this work, with their coverage of language pairs. Metrics that are not marked as reference-based are reference-free (a.k.a quality estimation) metrics.

cosine similarity of the MT output and reference based on Llama 2 sentence embeddings.

GEMBA-MQM (Kocmi and Federmann, 2023) uses three-shot prompting on the GPT-4 model with a single prompt and no language specific example.

KG-BERTScore (Wu et al., 2023) is the linear combination of KGScore and COMET-QE based BERTScore, where KGScore is incorporating multilingual knowledge graph into BERTScore.

MaTESe (Perrella et al., 2022) trains Deberta (for English) and InfoXLM (for German and Russian) encoders to identify MQM error spans and severity using WMT22 Metrics shared task MQM data.

mbr-metricx-qe (Naskar et al., 2023) uses the underlying technique of minimum bayes risks (MBR) decoding to develop a quality estimation metric. It uses an evaluator machine translation system and a reference-based utility metric (specifically BLEURT and MetricX) to calculate a quality estimation score of a model.

MEE4 (Mukherjee and Shrivastava, 2023) is an unsupervised, reference-based metric that is a weighted combination of syntactic similarity based on a modified BLEU score, lexical, morphological and semantic similarity using unigram matching and contextual similarity with sentence similarity scores from multilingual BERT.

MetricX-23 (Juraska et al., 2023) is a regression metric that finetunes the mT5-XXL checkpoint using direct assessment data from 2015-2020 and MQM data from 2020 to 2021 as well as synthetic data. **MetricX-23-QE** is the quality estimation variant that uses the source, instead of the reference, for scoring.

mre-labse-regular (Viskov et al., 2023) is a trained metric that is based on the encoder part of mT0-large model and contextual embeddings from LaBSE. It concatenates the source, reference and MT output as input.

XCOMET-Ensemble (Guerreiro et al., 2023) is an ensemble of a XCOMET-XL and two XCOMET-XXL checkpoints that result from the different training stages. XCOMET is similar to COMET but is trained for both regression and sequence tagging for identifying MQM error spans, where the intent is to make it a more interpretable learnt metric. **XCOMET-QE-Ensemble** is the quality estimation version.

XLsim (Mukherjee and Shrivastava, 2023) is a supervised reference-based metric that regresses on human scores provided by WMT (2017-2022) based on XLM-RoBERTa using a Siamese network architecture with CosineSimilarityLoss.

G Additional Figures

Here we show additional figures, including the full set of histograms for EN→DE (Figure 3), EN→HE (Figure 4), HE→EN (Figure 5) and ZH→EN (Figure 6) as well as the system scores for EN→DE (Figure 7), EN→HE (Figure 8), and ZH→EN (Figure 9).

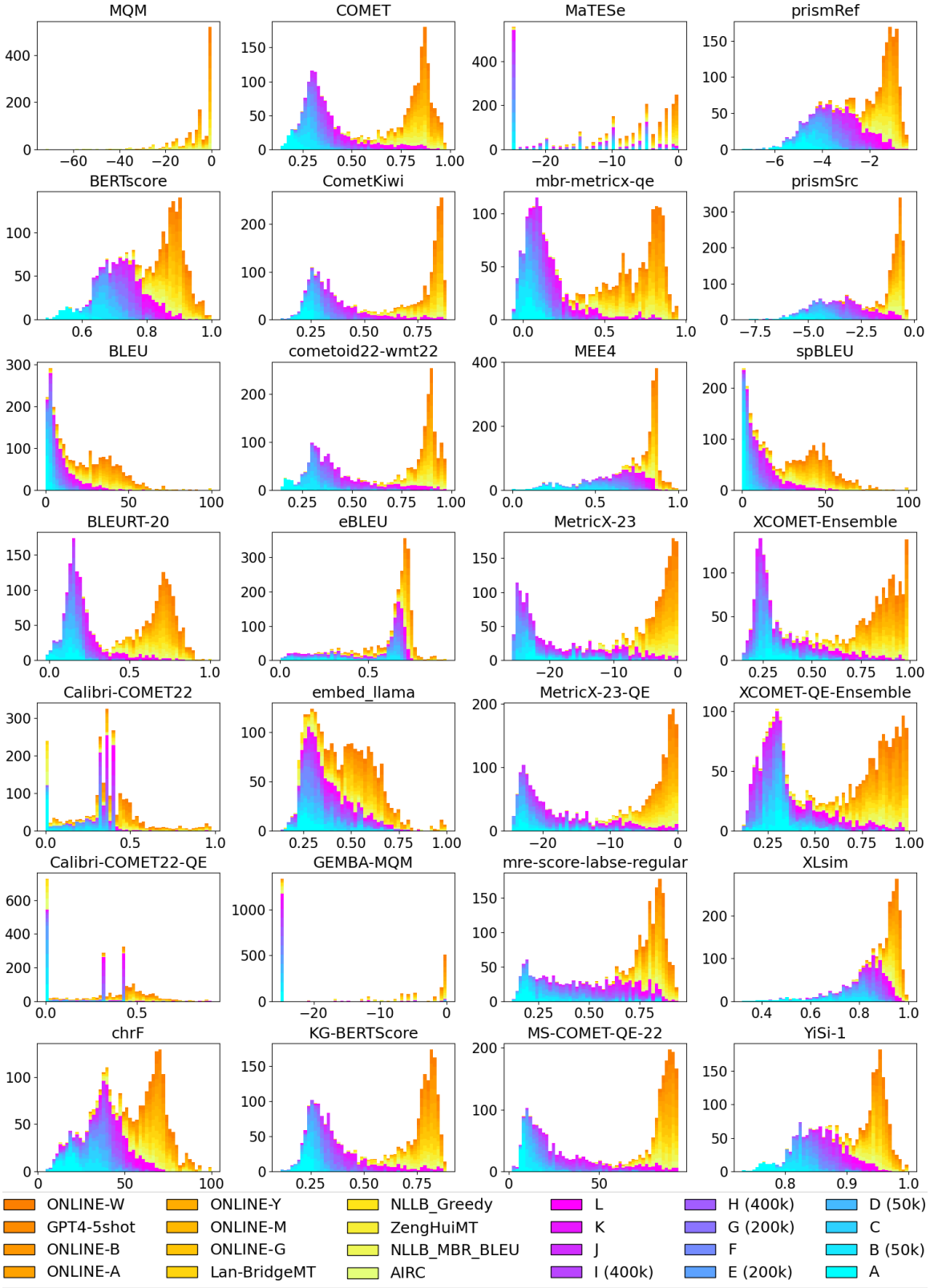


Figure 3: Stacked histograms (one subplot per metric) of segment scores for EN→DE across the challenge set (cool colours/bottom of the stacked histograms) and submitted WMT systems (warm colours/top of the stacked histograms).

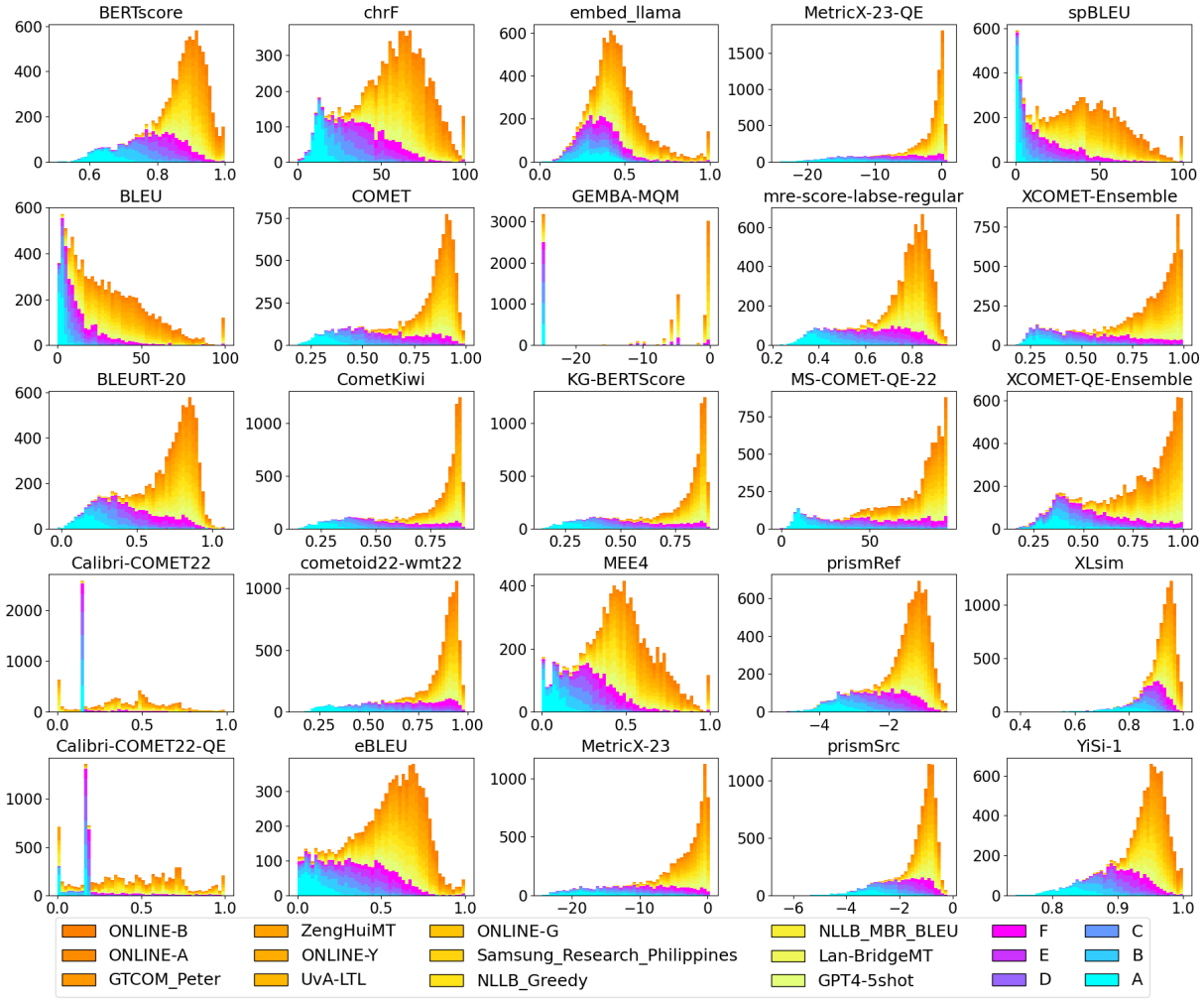


Figure 4: Stacked histograms of segment scores for EN→HE across the challenge set (cool colours/bottom) and submitted WMT systems (warm colours/top).

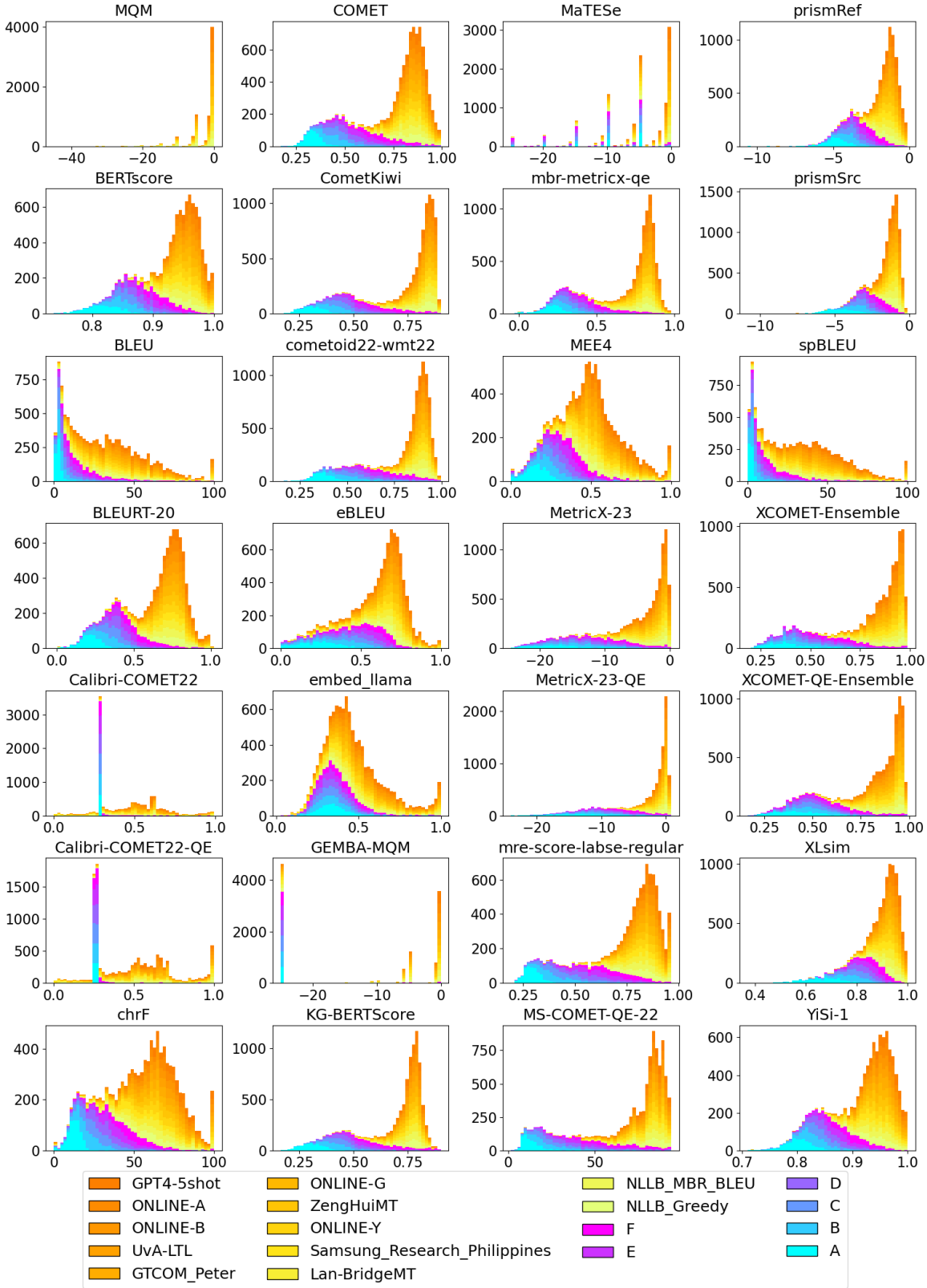


Figure 5: Stacked histograms of segment scores for HE→EN across the challenge set (cool colours/bottom) and submitted WMT systems (warm colours/top).

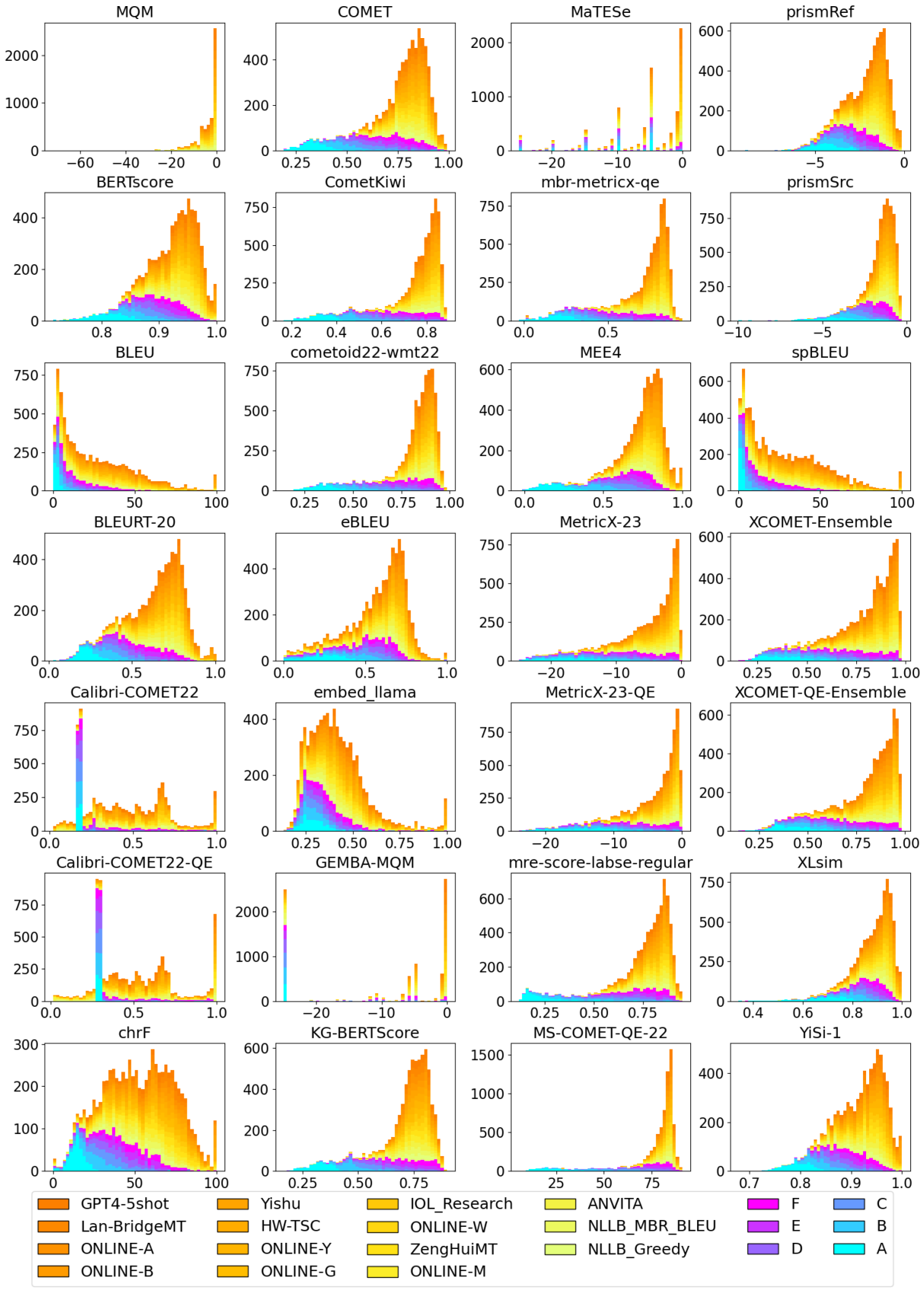


Figure 6: Stacked histograms of segment scores for ZH→EN across the challenge set (cool colours/bottom) and submitted WMT systems (warm colours/top).

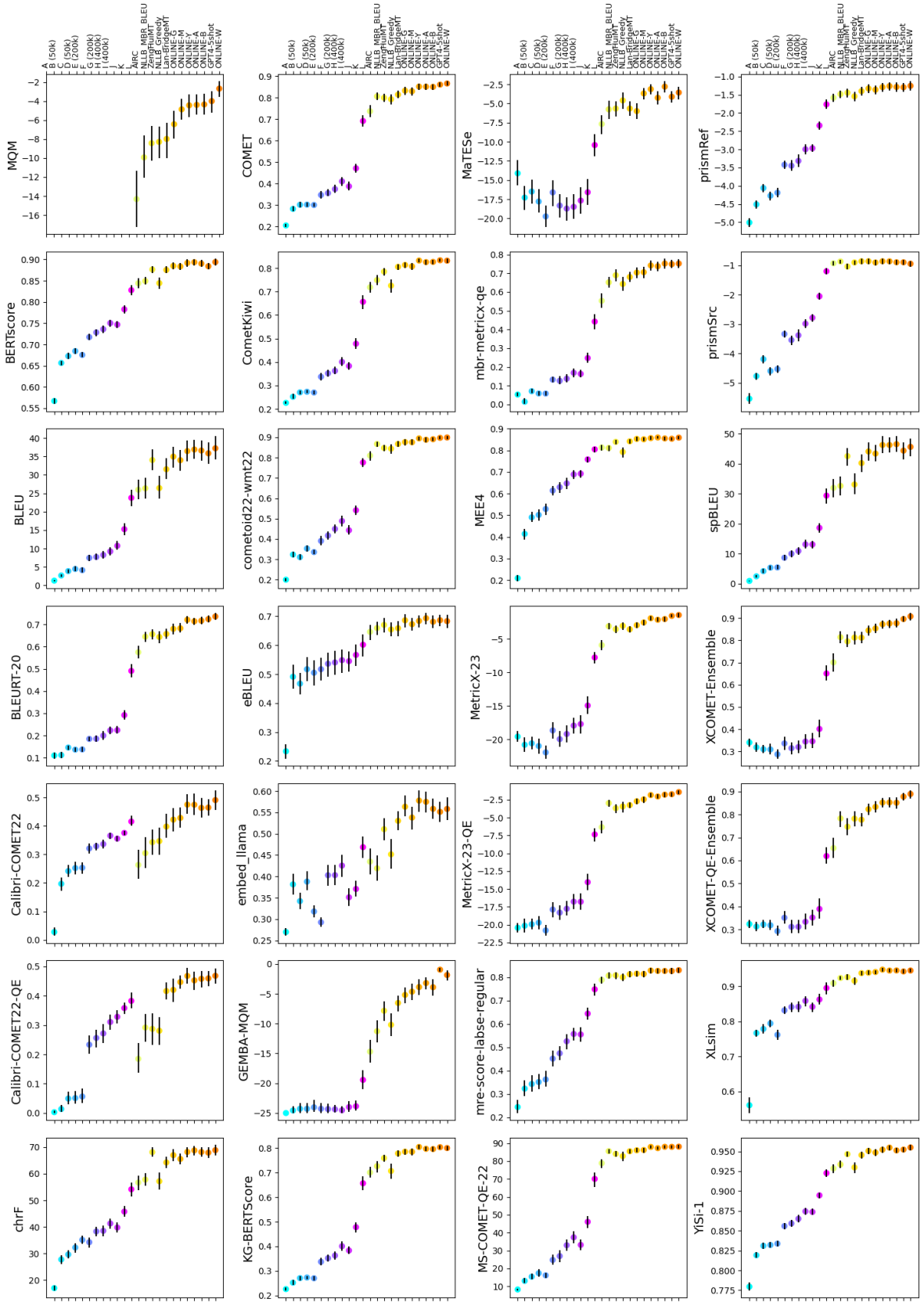


Figure 7: System average scores (with error bars computed via bootstrap resampling 1000 times for $p < 0.05$) for EN→DE across the challenge set (cool colours/left) and submitted WMT systems (warm colours/right). Our challenge set systems are ordered from left to right with BLEU scores, while the submitted WMT systems are ordered by MQM score on the news domain.

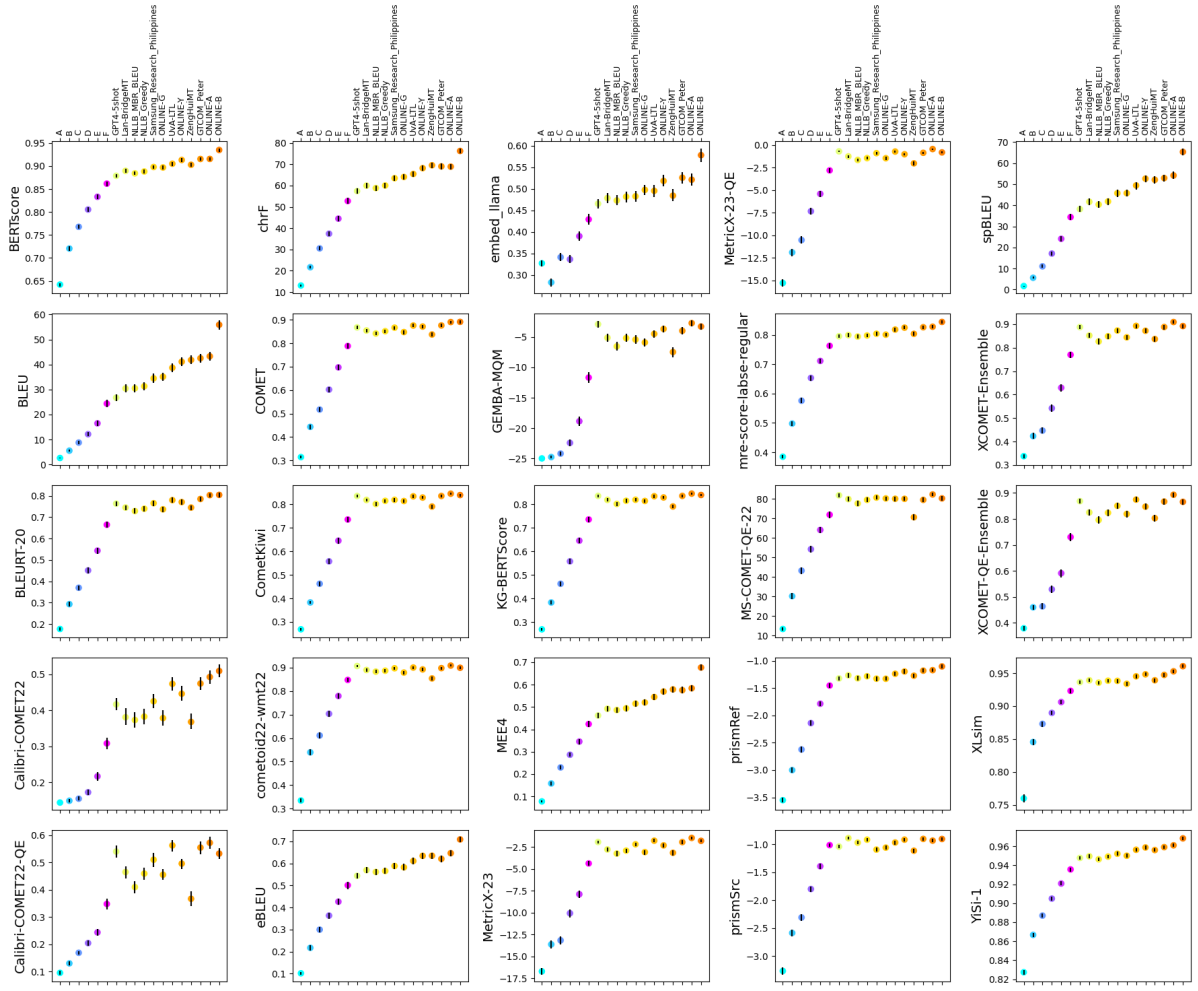


Figure 8: System average scores (with error bars computed via bootstrap resampling 1000 times for $p < 0.05$) for EN→HE across the challenge set (cool colours/left) and submitted WMT systems (warm colours/right). All systems are ordered from left to right by BLEU scores (as direct assessment scores were not yet available for EN→HE).

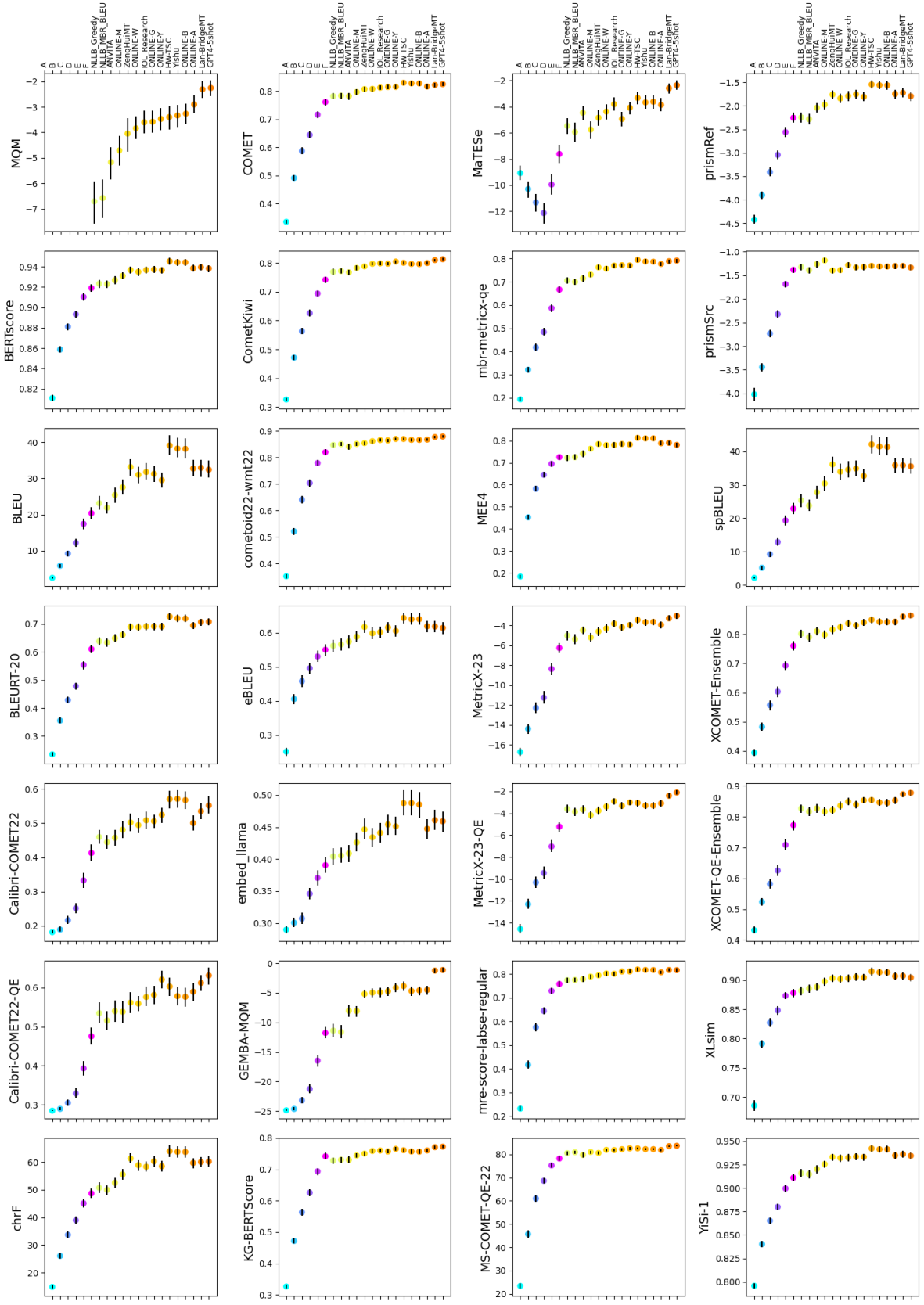


Figure 9: System average scores (with error bars computed via bootstrap resampling 1000 times for $p < 0.05$) for ZH→EN across the challenge set (cool colours/left) and submitted WMT systems (warm colours/right). Our challenge set systems are ordered from left to right with BLEU scores, while the submitted WMT systems are ordered by MQM score on the news domain.

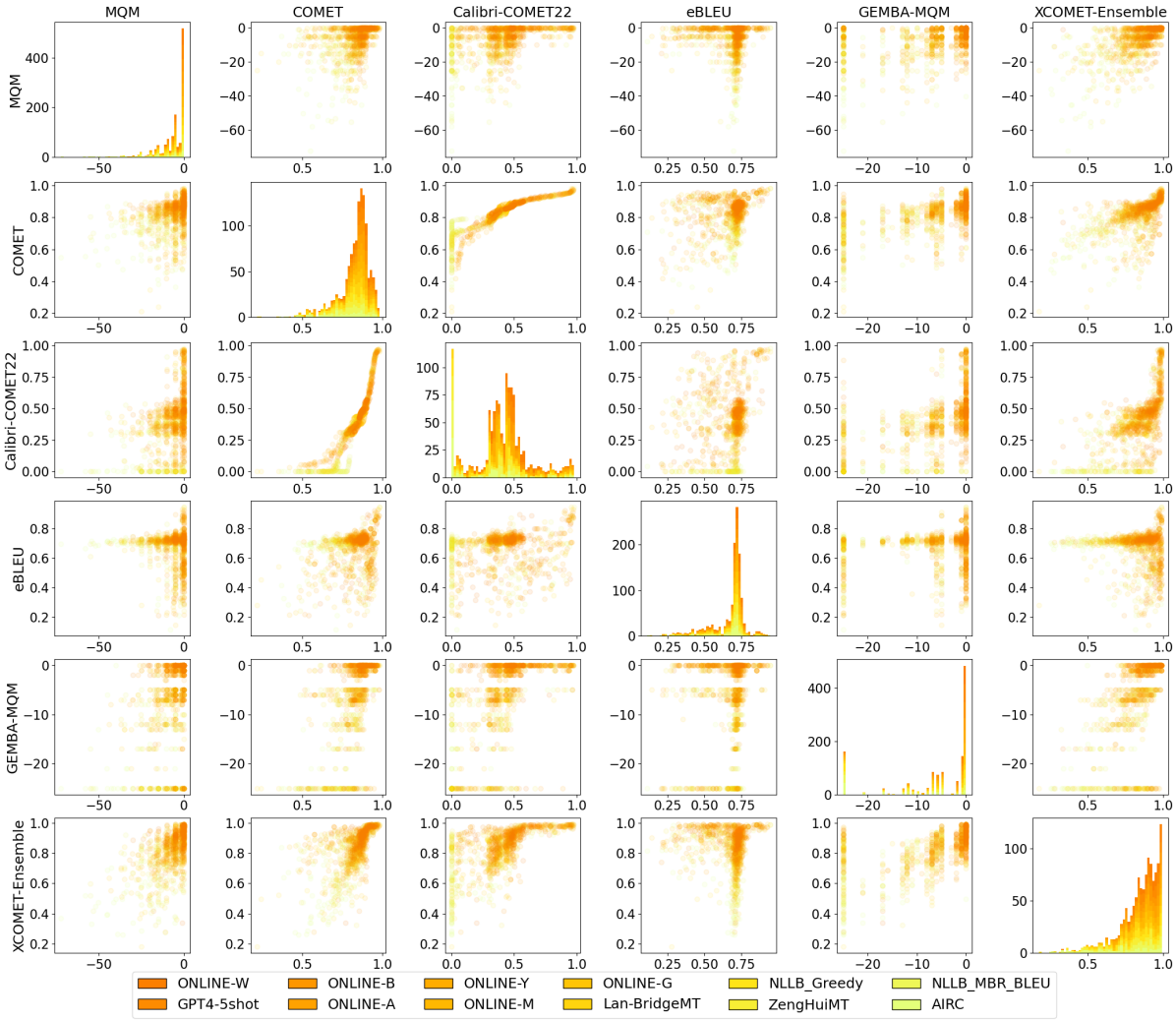


Figure 10: A subset of the metrics (and MQM scores) for EN→DE, showing only the high-quality WMT MT system submissions. The diagonal entries show stacked histograms of segment scores. The off-diagonal entries are scatterplots where each point is a single segment positioned according to the score assigned to it by row and column metrics; each point is coloured according to the MT system that produced it.

MEE4 and XLsim : IIIT HYD's Submissions' for WMT23 Metrics Shared Task

Ananya Mukherjee and **Manish Shrivastava**
Machine Translation - Natural Language Processing Lab
Language Technologies Research Centre
International Institute of Information Technology - Hyderabad
ananya.mukherjee@research.iiit.ac.in
m.shrivastava@iiit.ac.in

Abstract

This paper presents our contributions to the WMT2023 shared metrics task, consisting of two distinct evaluation approaches: a) **Unsupervised Metric** (MEE4) and b) **Supervised Metric** (XLSim). MEE4 represents an unsupervised, reference-based assessment metric that quantifies linguistic features, encompassing lexical, syntactic, semantic, morphological, and contextual similarities, leveraging embeddings. In contrast, XLsim is a supervised reference-based evaluation metric, employing a Siamese Architecture, which regresses on Direct Assessments (DA) from previous WMT News Translation shared tasks from 2017-2022. XLsim is trained using XLM-RoBERTa (base) on English-German reference and mt pairs with human scores. Here are the links for MEE4 [1] and XLsim[2] metrics.

1 Introduction

In recent times, there has been a growing interest in Neural Machine Translation (NMT) systems, leading to significant improvements in machine translation (MT) quality. Over the past few years, the field of MT evaluation has seen substantial advancements. Each year, the WMT conference hosts a metrics-shared task, where new evaluation metrics are introduced and those demonstrating a strong correlation with human judgments are highlighted from the array of newly devised metrics. In the last three years of the WMT Metrics Task (Freitag et al., 2022, 2021; Mathur et al., 2020), neural-based metrics have predominantly taken the lead. Nevertheless, n-gram-based and lexical-based metrics (Papineni et al., 2002; Popović, 2015) continue to be favored as automatic MT evaluation tools due to their flexibility and efficiency.

As a result, this year we participated in the metrics shared task, evaluating machine translation outputs using two types of metrics: an unsupervised metric and a supervised metric.

Unsupervised Metric: Our unsupervised metric, MEE4 (Mukherjee and Shrivastava, 2022), relies on a combination of lexical and embedding similarity measures. Notably, MEE4 demonstrated strong performance in the previous year's shared task (Freitag et al., 2022), surpassing several baseline metrics such as BERTscore (Zhang* et al., 2020), BLEU (Papineni et al., 2002), F101SPBLEU (Goyal et al., 2022) , and CHRf (Popović, 2015). In our efforts to improve its performance further this year, we conducted experiments with two different sentence embedding models: LaBSE (Feng et al., 2022) and the stsb-xlm-r-multilingual [3]. Interestingly, our findings indicated that MEE4, when equipped with LaBSE as the sentence embedding model, exhibited superior performance compared to the alternatives.

Supervised Metric: Unlike the existing neural models which are huge in size, our goal was to build a more compact supervised training model (XLsim) that offers improved performance. To achieve this, we created a SentenceTransformer model by combining a pre-trained transformer model with a pooling layer. This hybrid approach enables the generation of sentence embeddings, which can be compared using cosine similarity to assess similarity between sentences.

2 MEE4

MEE4 is an improved version of MEE focusing on computing contextual and syntactic equivalences, along with lexical, morphological, and semantic similarity. The goal is to comprehensively evaluate the fluency and adequacy of MT outputs while also considering the surrounding context. Fluency is determined by analyzing syntactic correlations, while context is evaluated by comparing sentence

[1] https://github.com/AnanyaCoder/WMT22Submission
[2] https://github.com/AnanyaCoder/XLsim
[3] https://huggingface.co/sentence-transformers/stsb-xlm-r-multilingual

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 800–805
December 6–7, 2023.

similarities using sentence embeddings. The ultimate score is derived from a weighted amalgamation of three distinct similarity measures: a) Syntactic similarity, which is established using a modified BLEU score. b) Lexical, morphological, and semantic similarity, quantified through explicit unigram matching. c) Contextual similarity, gauged by sentence similarity scores obtained from the Language-Agnostic BERT model (Feng et al., 2022).

In our experiments this year, we made adjustments to MEE4 while maintaining the same underlying architecture. Specifically, we computed the evaluation scores using a different sentence embedding model.

In addition to our previous choice, we utilized the stsb-xlm-r-multilingual model. This particular sentence-transformers model is designed to map sentences and paragraphs into a 768-dimensional dense vector space, making it suitable for various tasks such as clustering and semantic search. It's worth highlighting that the version of XLM-R (Conneau et al., 2020) we employed is considered a state-of-the-art model for multilingual Semantic Textual Similarity (STS) (Reimers and Gurevych, 2020).

2.1 Multilingual Sentence Encoders

Numerous multilingual sentence encoders, including mBERT (Devlin et al., 2018), consist of single self-attention networks. These models are pretrained on monolingual corpora in over 100 languages and are optimized for masked language modeling. Here, the model is tasked with predicting randomly selected tokens in the original text that have been replaced by a placeholder.

However, these pretrained multilingual sentence encoders often exhibit limited sensitivity to cross-language semantic similarity. To address this issue, Reimers and Gurevych employed human Semantic Textual Similarity (STS) annotations to enhance a pretrained multilingual sentence encoder, specifically BERT resulting in *stsb-xlm-r-multilingual* model .

In contrast, LaBSE differs slightly as it has been trained not only for masked language modeling but also for translation language modeling.

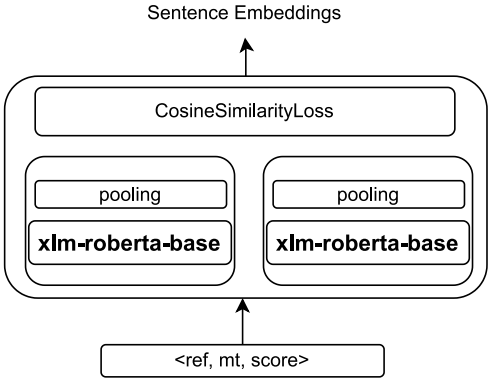


Figure 1: Illustration of our methodology using siamese network architecture
En: English IL- Indian Langauge

3 XLSim: MT Evaluation Metric based on Siamese Architecture

XLsim is a supervised reference-based metric that regresses on human scores provided by WMT (2017-2022). Using a cross-lingual language model XLM-RoBERTa-base[4] (Conneau et al., 2020), we train a supervised model using a Siamese network architecture with CosineSimilarityLoss.

3.1 Training Data

The WMT DA human evaluation data[5] (WMT17-WMT22) (Kocmi et al., 2022; Akhbardeh et al., 2021; Barrault et al., 2020, 2019; Bojar et al., 2018, 2017) contains raw score and z-score; we considered z-score for our training purpose by normalizing it to a range of 0-1.

3.2 Siamese Network Architecture

Similar to SBERT, we train the network with a Siamese Network Architecture (Reimers and Gurevych, 2019). In this siamese network, for each sentence pair, we pass *reference translation (ref)* and *hypothesis translation (mt)* through our network which yields the embeddings u und v. The similarity of these embeddings is computed using cosine similarity and the result is compared to the gold similarity score *(score)*. This allows our network to be fine-tuned and recognize sentence similarity. Figure 1 illustrates our XLsim training architecture.

While training, we used **CosineSimilarityLoss**, which automatically ensures training in a siamese network structure.

[4] https://huggingface.co/xlm-roberta-base
[5] https://huggingface.co/datasets/RicardoRei/wmt-da-human-evaluation

ref	I believe that financially, automakers are doing very well now, maintaining high sales margins.
mt	I believe car manufacturers are feeling very good financially right now, maintaining high sales margins.
score	0.77

Table 1: Input Example

3.3 CosineSimilarityLoss

CosineSimilarityLoss expects that the input consist of two texts and a float label. Refer Table 1.

It computes the vectors $u = model(input[0])$ and $v = model(input[1])$ and measures the cosine-similarity between the two. By default, it minimizes mean squared error loss.

3.4 Training Details

In our experiment, we focused on the en-de[6] language pair and utilized specific columns from the wmt-da-human-evaluation dataset, which included translation (mt), reference translation (ref), and z-score (score). Among the total 125,992 en-de samples available, we partitioned them as follows: 105,992 samples were used for training, 10,000 for validation, and another 10,000 for testing.

We employed a SentenceTransformer architecture to train our model, leveraging a multilingual pre-trained transformer model, XLM-RoBERTA base model. XLM-RoBERTa (Conneau et al., 2020) model is pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages.

We utilized the CosineSimilarityLoss function for a total of 4 training epochs. Our training setup involved a batch size 16, employing the Adam optimizer with a learning rate 2e-5 and a linear learning rate warm-up strategy over 10% of the training data. The entire training process was carried out on NVIDIA GPUs, specifically T4 x2.

3.5 Inference

To assess translation quality based on reference, our trained model generates embeddings for reference and translation sentences and subsequently calculates the cosine similarity between these embeddings. This similarity measure serves as a metric for evaluating the quality and similarity between the translation and reference text (refer figure 2).

[6]we chose the language-pair having a more significant number of samples than other language-pairs.

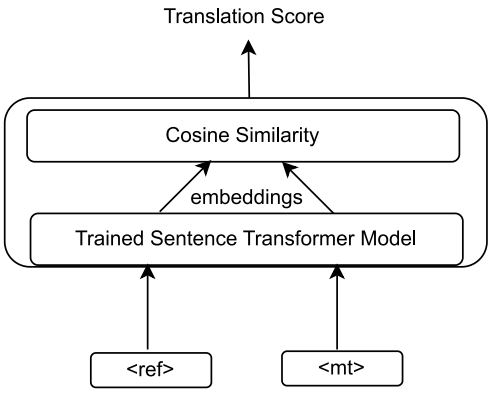


Figure 2: XLsim architecture at inference (to compute segment-level scores)

Model	COMET	XLsim
Size	2.32 GB	1.1 GB
Training Samples#	1,027,155	105,992
Pearson correlation	0.68	0.52

Table 2: Comparison with the SOTA neural metric based on Pearson Correlation with human scores.

Table 2 reports the comparison of our trained metric with the existing state-of-the-art metric, COMET (Rei et al., 2022) in terms of model size, total training samples and pearson correlation on the 10000 en-de samples (test samples see 3.4). It is worth noticing that the difference in correlation is 0.16 which is minute and model is 50% lesser in size.

4 WMT23 Metric Shared Task Submission

4.1 Segment Level Evaluation

For Segment-level task, we submitted the sentence-level scores obtained by our reference-based unsupervised metrics namely MEE4 (primary metric) and MEE4_stsb_xlm.

For the same Segment-level task, we also submitted the sentence-level scores obtained by our reference-based supervised evaluation metric (XLsim).

4.2 System Level Evaluation

To calculate the system-level score for each system, we take the average of the segment-level scores that we've derived. We employ a similar approach when computing system-level scores based on segment-level human annotations, such as DA's and MQM.

testset	lp	#sentences	XLsim	MEE4	MEE4_stsb_xlm
generaltest2023	en-de	6684	**0.67**	0.64	0.47
	zh-en	29640	0.68	**0.74**	0.59
	he-en	22920	0.76	**0.78**	0.62
challengeset	en-de	33470	**0.73**	0.71	0.64
	zh-en	6996	0.86	**0.91**	0.89
	he-en	9466	0.80	**0.86**	0.85

Table 3: Pearson correlation of evaluated scores on WMT23 submissions with COMET metric.

This suggests that a metric with a strong correlation at the segment level should also exhibit a robust correlation at the system level.

4.3 Results

Table 3 provides the details of the WMT23 Metric Shared Task test-set for the language pairs we investigated. However, it's important to note that the final and most comprehensive analysis will rely on the official results, where metric submissions are thoroughly compared to human judgments.

In our preliminary assessment, we have reported Pearson correlation scores for the submitted metrics when compared to COMET at the segment-level. This analysis helps us gauge the performance of the three metrics in relation to the state-of-the-art metric. In case of Unsupervised metrics, it appears that MEE4, which utilizes LaBSE, outperforms MEE4_stsb_xlm, which employs stsb-xlm-r-multilingual as its sentence embedding model. This difference in performance may be attributed to the training techniques applied to LaBSE, which involve both masked language modeling and translation language modeling, making it more effective for the task. Indeed, it's evident that XLsim exhibits a relatively strong correlation with COMET, almost exceeding 0.7. However, when compared to MEE4, there is a mild decrease in performance, particularly in the zh-en (Chinese to English) and he-en (Hebrew to English) language pairs, where the correlation drops by approximately 0.06.

This slight decline in performance for XLsim in certain language pairs could be attributed to the fact that even though XLsim utilizes the pre-trained multilingual XLM-Roberta model, the training data (ref, mt) was primarily in the German (de) language.

5 Conclusion and Future Work

In this paper, we describe our submissions to the WMT23 Metrics Shared Task. Our submission includes segment-level and system-level translation evaluation scores for sentences of three language pairs English-German (en-de), Chinese-English (zh-en) and Hebrew to English (he-en). We evaluate this year's test set using: a)two unsupervised metrics, *MEE4 and MEE4_stsb_xlm*. These metrics are based on lexical and embedding similarity match that evaluates the translation on various linguistic features (syntax,lexical, morphology, semantics and context) ; b) a supervised metric, XLsim that learns on en-de WMT DA human evaluation data from 2017-2022. It is observed that all the three metrics displayed a positive correlation (>0.5) with the baseline metric COMET.

Certainly, there are promising research directions to explore, especially in the realm of metric enhancement. In our future work, we intend to delve deeper into these areas:

MEE4 Metric Improvement: One of our primary objectives is to refine and enhance MEE4, seeking more efficient approaches that can better estimate translation quality while achieving higher agreement with human judgments. This might involve exploring novel techniques in sentence embedding, fine-tuning, or leveraging additional linguistic information.

XLsim Enhancement: For XLsim, we plan to boost its performance by optimizing the training data. This involves ensuring that it is trained on a more diverse set of languages and data to improve its cross-lingual capabilities. Simultaneously, we aim to maintain its compactness and ensure it remains trainable with fewer computational requirements.

These future research directions hold the potential to contribute significantly to the field of machine translation evaluation, ultimately leading to more robust and accurate metrics that align closely with human assessments.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Ananya Mukherjee and Manish Shrivastava. 2022. Unsupervised embedding-based metric for MT evaluation with improved human correlation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 558–563, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Quality Estimation using Minimum Bayes Risk

Subhajit Naskar, Daniel Deutsch, and Markus Freitag
Google
{snaskar,dandeutsch,freitag}@google.com

Abstract

This report describes the Minimum Bayes Risk Quality Estimation (MBR-QE) submission to the Workshop on Machine Translation's 2023 Metrics Shared Task. MBR decoding with neural utility metrics like BLEURT is known to be effective in generating high quality machine translations. We use the underlying technique of MBR decoding and develop an MBR based reference-free (quality estimation) metric. Our method uses an evaluator machine translation system and a reference-based utility metric (specifically BLEURT and MetricX) to calculate a quality estimation score of a model's output. We report results related to comparing different MBR configurations and utility metrics.

1 Introduction

The task of quality estimation (QE) is to assign a sentence- or word-level quality score to a machine translation (MT) output without the use of a reference translation. In this paper, we describe the methodology used in our sentence-level QE metric submission to the 2023 Workshop on Machine Translation's Metrics Shared Task.

Minimum Bayes Risk (MBR) decoding has been widely used in machine translation to address the limitation of MAP decoding (Kumar and Byrne, 2004; Eikema and Aziz, 2020; Müller and Sennrich, 2021). Freitag et al. (2021b) showed applying MBR decoding using BLEURT (Sellam et al., 2020) as a utility function can out-perform beam search decoding.

MBR decoding can be viewed as a method for reranking candidate outputs from an MT system. It first samples a set of hypothesis translations from the model, scores each hypothesis against a set of pseudo-references (generally, the same set of sample hypotheses) with a utility metric, then selects the hypothesis with the highest average score to be the final translation.

Central to MBR is assigning a quality score to a hypothesis translation without the use of a reference. Because this decoding procedure has been successful in improving the quality of translations from an MT system, in this work, we explore how MBR could be repurposed as a QE metric.

Our proposed metric uses an MT system in conjunction with a utility metric to assign a quality score to a translation without using a reference. The metric assigns a score to a hypothesis translation by using the utility metric to evaluate the hypothesis against a set of pseudo-references that are sampled from the MT model.

In this work, we experiment with creating a metric that uses different MT systems, utility functions, and different pseudo-reference pool sizes. Our experiments demonstrate that (1) a better utility function results in better MBR-QE scores, (2) the choice of MT system can have significant impact on QE metric performance, and (3) the size of the pseudo-reference pool does not have a significant impact on overall metric quality.

Based on our experiments, we chose our primary MBR-QE submission to be an in-house encoder-decoder model with MetricX (Freitag et al., 2022) as the utility function with a pseudo-reference pool size of 256.

2 Metric Descriptions

MBR decoding has two components: an MT system and a utility function. The MT model $P_{model}(y|x)$ estimates the probability of target segments y given a source segment x. The utility function estimates the quality of a translation h given a reference translation r. The best hypothesis is selected using the expected utility with respect to a finite sample generated by the model. The underlying assumption is that the model provides good approximation for the true distribution of human translations.

We adopt that assumption to develop an MBR-

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 806–811
December 6–7, 2023.

based quality estimation metric. The MBR-QE metric uses an MT system P to generate the set of pseudo-references, denoted $\hat{R}$. Then, the utility function defines the quality of a translation h given $\hat{R}$ as the average score over all of $\hat{R}$:

$$\text{MBR-QE}(h) = \frac{1}{|\hat{R}|} \sum_{\hat{r} \epsilon \hat{R}} u(h, \hat{r}) \quad (1)$$

This methodology has multiple potential pitfalls. First, because the distribution P is used to substitute for the distribution of human translations, any significant divergence between these two distributions will lead to the QE score becoming inconsistent because the pool of pseudo-references will not resemble human references. This can be mitigated by using a high quality MT system. Arguably, the MT system should have better performance compared to the MT models that are being evaluated.

Second, our QE metric is dependent on the quality of the utility function. If it has limitations or biases, they will affect the predicted quality scores and introduce inconsistencies between the QE score and ground-truth human quality scores.

We next discuss the experimental setup for analyzing our proposed QE metric.

3 Experimental Setup

3.1 Pseudo-Reference Generation

Our MBR-QE metric relies on the assumption that if the MT system that generates the pseudo-references can be used as an approximation for the distribution of human translations, then the aggregated utility metric score can be used a quality estimate for hypothesis. Therefore, the MT model and method for generating pseudo-references is critical for the effectiveness of this metric.

MT Systems. The MT system used for our shared task submission is an in-house encoder-decoder translation model that is similar to the Google Translate production model. In this report, we also experiment with generating pseudo-references from the PaLM 2 (Bison) large language model (Anil et al., 2023) using 5-shot prompting.

Sampling Method. We generate pseudo-references from the MT system using epsilon sampling (Hewitt et al., 2022; Freitag et al., 2023) with $p = 0.02$ and sampling temperature 1.0. We experiment with using a different number of pseudo-references.

3.2 Utility Functions

Freitag et al. (2021b) showed that MBR decoding works well with neural evaluation metrics. We experiment with 2 neural metrics as the utility function in MBR-QE.

BLEURT v0.2 (Sellam et al., 2020; Pu et al., 2021): BLEURT v0.2 is a learned regression-based metric that is trained to predict the quality of a translation given a reference. It is pre-initialized with RemBERT (Chung et al., 2020) and finetuned using a combination of WMT human evaluation data from 2015-2019 and synthetic data.

MetricX (Freitag et al., 2022): MetricX is a learned regression-based metric that is based on mT5 (Xue et al., 2021). It is trained on a combination of direct assessment and MQM (Lommel et al., 2014; Freitag et al., 2021a) data that was collected by WMT. We use the reference-based version that uses mT5-XXL.

3.3 Meta-Evaluation

We use four different correlations to calculate the metrics' agreements with human judgments. At the system-level, we use pairwise accuracy (Kocmi et al., 2021) and Pearson's r. System-level Pearson's r captures how strong the linear relationship is between the metric and human scores for MT systems. Pairwise accuracy evaluates a metric's ranking of MT systems by calculating the proportion of all possible pairs of MT systems that are ranked the same by the metric and human scores.

At the segment-level, we use group-by-item pairwise accuracy with tie calibration (Deutsch et al., 2023) and no-grouping Pearson's r. The no-grouping Pearson's r calculates the linear relationship between the metric and human scores across translations from every system and document. The group-by-item pairwise accuracy calculates the proportion of all possible pairs of translations for the same input segment that are ranked the same or tied by the metric and human. Then the accuracy scores are averaged over all possible input segments. We use tie calibration (Deutsch et al., 2023) that automatically introduces ties into metric scores based on a threshold. This tie calibration is required as regression-based metrics rarely predict ties.

Our experiments are performed using the WMT'22 English-to-German (en-de) and Chinese-to-English (zh-en) MQM ratings (Freitag et al., 2022). These datasets are commonly used for meta-

evaluation and are the latest available from the Metrics Shared Task. We did not evaluate using en-ru since it is not included as a language pair in the WMT'23 evaluation.

4 Experimental Results

The main experimental results are shown in Tables 1 and 2. Table 1 compares the two utility functions with various pseudo-reference pool sizes when using the in-house MT system, and Table 2 does the same but for the PaLM 2-based system.

Comparing Utility Functions. For both MT systems and all pseudo-reference pool sizes, the MBR-QE metric that uses MetricX as a utility function in general has higher correlations than when BLEURT is used. This result is expected since MetricX was the best performing metric in the WMT'22 evaluation. This is evidence that the quality of the utility function is important for the quality of the MBR-QE score.

Comparing MT Systems. When comparing whether the encoder-decoder MT system or PaLM 2 is used to generate the pool of pseudo-references, there is no clear winner between the two. The MBR-QE score has a higher correlation at the segment-level with the encoder-decoder model, but the correlations are higher at the system-level with PaLM 2. It is not clear why this is the case.

Pseudo-Reference Pool Size. Overall, the correlations are surprisingly stable for each of the different numbers of pseudo-references. Most of the differences comes between pairwise accuracy at the system-level, but this correlation can be sensitive; there are not many system pairs, so if one or two system rankings change, it can have a large impact on the overall accuracy. In the future, we could explore decreasing the pseudo-reference pool size even further to understand its impact on the overall MBR-QE metric quality.

Comparing to Other Metrics. Table 3 shows the comparion between our submission, denoted MBR-QE, to other QE metrics that were the top-performing QE metrics in the WMT'22 Metrics Shared Task, CometKiwi (Kepler et al., 2019; Rei et al., 2022b) and UniTE-src (Wan et al., 2022). The table additionally contains results for the best reference-based metrics MetricX and COMET-22 (Rei et al., 2022a).

Compared to the QE metrics, MBR-QE in general has the best-performance across most evaluation settings, demonstrating that it is a state-of-the-art QE metric. In some cases, it even out-performs the reference-based metrics, namely in the system-level Pearson correlation.

MBR-QE leverages MetricX as the utility function. MBR-QE still under-performs with respect to MetricX, demonstrating that the human references are still valuable and that the pseudo-references do not perfectly match the distribution of human translations, which is expected given that the MT system is not perfect. However, the gap in performance between the two metrics is relatively small in some settings.

4.1 Submission Summary

Both of our submissions to the Metrics Shared task use the in-house MT system to generate 256 pseudo-references with epsilon sampling ($p = 0.02$ and temperature 1.0). Our primary submission uses MetricX as the utility function, and the contrastive submission uses BLEURT.

5 Related Work

Incorporating evaluation metrics into reranking the outputs from MT systems has been very successful. For example, the Freitag et al. (2021b) showed that reranking translations with BLEURT as part of MBR produced higher-quality translations. This work served as the inspiration for our QE metric submission.

Research on quality estimation focuses on predicting word- and sentence-level quality scores (Zerva et al., 2022). The most successful approaches to predicting sentence-level scores are learned regression-based metrics that are trained to predict ground-truth quality scores, like CometKiwi (Kepler et al., 2019; Rei et al., 2022b) or UniTE-src (Wan et al., 2022). Our metric is quite different from these approaches in that it is not directly trained to predict quality scores, but rather it leverages a reference-based metric combined with an MT system to score a translation. To the best of our knowledge, ours is the first metric that uses MBR to build a QE metric.

6 Conclusion

In this report, we proposed a new QE metric called MBR-QE that repurposes an MT system in combination with MBR to score a translation without ac-

Utility Metric	Pseudo-Ref Pool Size	SEG pairwise acc. en-de	SEG pairwise acc. zh-en	SEG Pearson en-de	SEG Pearson zh-en	SYS pairwise acc. en-de	SYS pairwise acc. zh-en	SYS Pearson en-de	SYS Pearson zh-en
BLEURT	64	0.5772	0.5142	0.4750	0.4628	0.7692	0.7802	0.6773	0.8907
	128	0.5777	0.5145	0.4752	**0.4631**	0.7692	0.7802	0.6749	0.8899
	256	0.5782	0.5151	0.4747	0.4626	0.7692	0.7692	0.6751	0.8901
MetricX	64	**0.5986**	0.5292	0.4891	0.4513	0.7564	**0.8132**	**0.8654**	0.8654
	128	0.5944	0.5300	0.4873	0.4519	**0.7821**	**0.8132**	0.8391	0.9579
	256	0.5979	**0.5306**	**0.4897**	0.4524	0.7692	**0.8132**	0.8647	**0.9586**

Table 1: MBR-QE correlations on the WMT'22 MQM data comparing when BLEURT and MetricX are used as utility functions with different pseudo-reference pool sizes are sampled from the in-house encoder-decoder model.

Utility Metric	Pseudo-Ref Pool Size	SEG pairwise acc. en-de	SEG pairwise acc. zh-en	SEG Pearson en-de	SEG Pearson zh-en	SYS pairwise acc. en-de	SYS pairwise acc. zh-en	SYS Pearson en-de	SYS Pearson zh-en
BLEURT	64	0.5614	0.4824	0.4261	0.4153	0.8077	0.7802	0.7636	0.9253
	128	0.5612	0.4827	0.4246	0.4143	0.8077	0.7802	0.7631	0.9250
	256	0.5616	0.4831	0.4249	0.4152	0.8077	0.7692	0.7635	0.9249
MetricX	64	0.5764	**0.5022**	0.4574	0.4259	0.7949	**0.8571**	**0.9154**	**0.9846**
	128	0.5737	0.5000	0.4621	0.4339	**0.8333**	0.8242	0.9145	0.9844
	256	**0.5767**	0.5021	**0.4626**	**0.4265**	0.8077	**0.8571**	0.9212	0.9845

Table 2: MBR-QE correlations on the WMT'22 MQM data comparing when BLEURT and MetricX are used as utility functions with different pseudo-reference pool sizes are sampled from using PaLM 2 as a translation system with 5-shot propmting.

Metric	SEG pairwise acc. en-de	SEG pairwise acc. zh-en	SEG Pearson en-de	SEG Pearson zh-en	SYS pairwise acc. en-de	SYS pairwise acc. zh-en	SYS Pearson en-de	SYS Pearson zh-en
Quality Estimation (Reference-Free) Metrics								
MBR-QE	**0.598**	**0.531**	**0.490**	0.452	**0.769**	**0.813**	**0.865**	**0.959**
COMETKIWI	0.572	0.509	0.432	**0.509**	0.692	0.758	0.674	0.866
UNITE-SRC	0.582	0.508	0.397	0.404	0.742	0.708	0.509	0.874
Reference-Based Metrics								
MetricX	0.605	0.544	0.549	0.581	0.829	0.867	0.847	0.920
COMET-22	0.594	0.536	0.512	0.585	0.790	0.886	0.771	0.942

Table 3: A comparison of our submission, denoted MBR-QE (scoring translations with MetricX against translations generated by our in-house MT system) to other QE metrics (top) and reference-based metrics (bottom). MBR-QE is overall the best-performing metric amongst the QE metrics, and it even improves over the reference-based metrics in system-level Pearson.

cess to a reference. Our experiments demonstrated that the choice of MBR utility function is important, the choice of MT system can impact downstream metric correlations, and the pseudo-reference pool size does not have a significant impact on results.

References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. PaLM 2 Technical Report.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models. *arXiv preprint arXiv:2010.12821*.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties Matter: Modifying Kendall's Tau for Modern Metric Meta-Evaluation.

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. 2023. Epsilon sampling rocks: Investigating sampling strategies for minimum bayes risk decoding for machine translation.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2021b. Minimum bayes risk decoding with neural metrics of translation quality. *CoRR*, abs/2111.09388.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chikiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

John Hewitt, Christopher Manning, and Percy Liang. 2022. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3414–3427, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019. OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122, Florence, Italy. Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica*, (12):0455–463.

Mathias Müller and Rico Sennrich. 2021. Understanding the properties of minimum Bayes risk decoding in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 259–272, Online. Association for Computational Linguistics.

Amy Pu, Hyung Won Chung, Ankur P Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for mt. In *Proceedings of EMNLP*.

Ricardo Rei, JosÃ© G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and AndrÃ© F. T. Martins. 2022a. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 578–585, Abu Dhabi. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Thibault Sellam, Amy Pu, Hyung Won Chung, Sebastian Gehrmann, Qijun Tan, Markus Freitag, Dipanjan Das, and Ankur Parikh. 2020. Learning to evaluate translation beyond English: BLEURT submissions to the WMT metrics 2020 shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 921–927, Online. Association for Computational Linguistics.

Yu Wan, Keqin Bao, Dayiheng Liu, Baosong Yang, Derek F. Wong, Lidia S. Chao, Wenqiang Lei, and Jun Xie. 2022. Alibaba-translate china's submission for wmt2022 metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 586–592, Abu Dhabi. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

The SLIDE metric submission to the WMT 2023 metrics task

Vikas Raunak and **Tom Kocmi** and **Matt Post**
Microsoft
Redmond, Washington, USA
{viraunak,tom.kocmi,mattpost}@microsoft.com

Abstract

We describe our submission of a new metric, SLIDE (Raunak et al., 2023), to the WMT 2023 metrics task. SLIDE is a reference-free quality-estimation metric that works by constructing a fixed sentence-length window over the documents in a test set, concatenating chunks and then sending them for scoring as a single unit by COMET (Rei et al., 2022a). We find that SLIDE improves dramatically over its context-less counterpart on the two WMT22 evaluation campaigns (MQM and DA+SQM).

1 Introduction

Reference-based metrics such as COMET typically perform far above their source-based quality estimation (QE) counterparts. One explanation is that the human reference provides an answer or grounding to many types of translation ambiguities, such as pronoun selection, that may be impossible to predict from just a single input sentence alone. A handful of approaches have looked at extending metrics with source- and target-side context (Vernikos et al., 2022; Deutsch et al., 2023; Raunak et al., 2023) in hopes of providing stronger correlation with human judgments. We base our submission on SLIDE (Raunak et al., 2023), which explicitly postulates and provides evidence for the claim that source-side context may work to provide the same information as human references.

2 Metric settings

SLIDE is parameterized by (w, s), a window and a stride. The window, w, is a fixed-size sentence window that is moved across each document in a test set. The sentences in the window are concatenated on the source and system systems with a space, and then sent directly to the underlying QE model, COMETKiwi (Rei et al., 2022b) in our submission, for evaluation as a single chunk. The window is then incremented by s sentences, and a new value computed. These values are treated independently, summed and averaged over a test set in typical fashion. Documents that are shorter than the window size, and the "remainder" portions of documents that cannot be perfectly tiled by the window and stride, are skipped.

In practice, we used a (w, s) value of $(6, 6)$ for all languages except EN-DE and DE-EN. For those languages, the data was provided at the paragraph level. We therefore simply took the provided segmentations one-by-one, without providing a window or stride. We chose this value because it had some of the best reported results in Raunak et al. (2023, Figure 1). Table 1 repeats Table 2 from

Metric	MQM	DA+SQM
® metricx_xl_DA_2019	0.865	0.850
® metricx_xxl_MQM_2020	0.850	0.861
® BLEURT-20	0.847	0.827
® metricx_xl_MQM_2020	0.843	0.859
SLIDE$(6, 6)$	0.843	0.838
® COMET-22	0.839	0.839
® COMET-20	0.836	0.823
® Doc-COMET	0.836	0.810
® UniTE	0.828	0.847
® MS-COMET-22	0.828	0.830
® UniTE-ref	0.818	0.838
® MATESE	0.810	-
® YiSi-1	0.792	0.782
COMETKiwi (WMT-22)	0.788	0.832
COMETKiwi (public)	0.770	0.816
Doc-COMET	0.752	0.810
® chrF	0.734	0.758
® BLEU	0.708	0.704

Table 1: Pairwise system accuracy against the WMT22-MQM and DA+SQM annotations. Metrics that use a reference are marked with ®. We mark our entries in bold. **COMETKiwi (public)** uses no context. Our entry to the WMT23 task, SLIDE (6,6), improves over it in both settings.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 812–814
December 6–7, 2023.

their paper, depicting results on the WMT22 tasks with the pairwise accuracy (Kocmi et al., 2021). Our entries are marked in bold. SLIDE improves dramatically over its context-less counterpart. We also call attention to COMETKiwi (WMT-22); this is the number from the official submission to the task, which performs much better than the publicly available model.

3 Results

The WMT23 test set (Freitag et al., 2023; Kocmi et al., 2023) for each language pair comprises a set of documents containing between 1 and 173 lines, with a mean of 9.7 and a median of 7 across 14 language pairs.

At the time of publication, official results were not available, so we cannot comment on how well the strong results from Raunak et al. (2023) generalized to the new settings in WMT23.

We note also that we discovered after the submission that a bug in our code resulted in debugging output appearing in the data to be scored by COMET. This unfortunately affects the scores and means that SLIDE's placement in the official rankings are incorrect.

4 Conclusion

In this system description, we presented our submission to the WMT 2023 metrics task. SLIDE is designed as a reference-free quality-estimation metric which leverages the strength of contextual information by constructing a fixed sentence-length window over documents in a test set. The initial findings from Raunak et al. (2023) showcased the potential of SLIDE to deliver enhanced performance over context-less metrics, particularly in the WMT22 evaluation campaigns.

While we anticipate the official results from the WMT23 metrics task, bug in our code might have affected SLIDE's standing in the rankings.

We believe that SLIDE is a step forward in our collective endeavor to create metrics that align more closely with human judgments. Future works may explore optimizing window and stride configurations or integrating advanced algorithms to further exploit the potential of context in quality estimation tasks.

References

Daniel Deutsch, Juraj Juraska, Mara Finkelstein, and Markus Freitag. 2023. Training and meta-evaluating machine translation evaluation metrics at the paragraph level.

Markus Freitag, Nitika Mathur, Chi kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frédéric Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of wmt23 metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and Maja Popović. 2023. Findings of the 2023 conference on machine translation (WMT23). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Vikas Raunak, Tom Kocmi, and Matt Post. 2023. Slide: Reference-free evaluation for machine translation using a sliding document window.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any

pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Semantically-Informed Regressive Encoder Score

Vasiliy Viskov[1]*, George Kokush[2]*, Daniil Larionov[3]*, Steffen Egger[3], and Alexander Panchenko[1,4]
[1]Skoltech, [2]HSE University, [3]NLLG Group, Bielefeld University, [4]AIRI
{vasiliy.viskov, a.panchenko}@skol.tech {daniil.larionov, steffen.eger}@uni-bielefeld.de

Abstract

Machine translation is a natural language generation (NLG) problem that involves translating source text from one language to another. Like every task in the machine learning domain, it requires an evaluation metric. The most obvious one is human evaluation; however, it is expensive, time-consuming, and not easily reproducible automatically. In recent years, with the introduction of pretrained transformer architectures and large language models (LLMs), state-of-the-art results in automatic machine translation evaluation have significantly improved in terms of correlation with expert assessments. We introduce MRE-Score, which stands for seMantically-informed Regression Encoder Score. It is an approach that constructs an automatic machine translation evaluation system based on a regression encoder and contrastive pretraining for the downstream problem.

1 Introduction

WMT Metrics Shared Task (Freitag et al., 2022) is a machine learning competition where participants have to construct an automatic evaluation system for machine translation for several language pairs. For WMT23 Metrics Shared Task[1], three language pairs are considered: English-German (en-de), Chinese-English (zh-en), and Hebrew-English (he-en). For each source sentence, there is a corresponding target machine-translated text and a reference human translation. The main goals of this competition are:

1. To achieve the strongest correlation with human judgment of translation quality over a diverse set of machine translation systems.

2. To illustrate the suitability of an automatic evaluation metric as a surrogate for human evaluation.

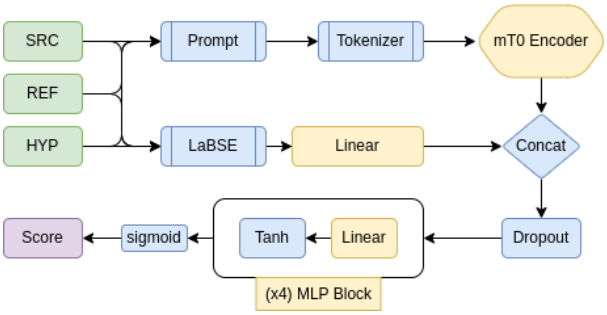


Figure 1: Final model architecture. Blocks in blue represents static components that were not trained. Blocks in yellow represent trained parts of the model.

3. To test the robustness of metrics when evaluating domains other than news data.

4. To create high-quality datasets for developing and evaluating metrics.

Within the WMT23 Metrics competition, our investigation focuses on the approach of constructing evaluation models to solve the regression problem based on expert degrees. Specifically, we construct regression models using a pretrained transformer encoder from the mT0 model family (Muennighoff et al., 2022), both vanilla models and with additional contrastive representation tuning. mT0 is finetuned version of mT5 (Xue et al., 2020), multilingual transformer model, which demonstrated capabilities of crosslingual generalization to unseen tasks and languages. Similar approaches demonstrated the best results in WMT21 (Freitag et al., 2021) and WMT22 (Freitag et al., 2022) Shared Tasks.

We release our code and pre-trained models openly to foster further research.[2]

2 Related Work

Evaluation metrics In NLG evaluation, one can differentiate four types of approaches for model

*Equal contribution
[1]https://wmt-metrics-task.github.io
[2]https://github.com/v-vskv-v/WMT23-MRE-Score

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 815–821
December 6–7, 2023.

construction: (1) classical lexical overlap models, on the one hand, and LLM-based approaches based on (2) unsupervised matching, (3) regression, and (4) zero-shot prompting, on the other hand. Classical lexical overlap methods measure overlap between source, reference, and target sentence n-grams (Papineni et al., 2002; Lin, 2004; Banerjee and Lavie, 2005). Modern unsupervised matching-based approaches use large language model (LLM)-based encoders such as BERT to compute the semantic similarity between reference and hypothesis texts (Zhang et al., 2019; Zhao et al., 2019) or between source and hypothesis texts (Zhao et al., 2020). In modern regression approaches, models are fine-tuned to predict human evaluation scores. Generally, they consist of a transformer encoder model and a regression head. As input, they can use late binding with source, reference, and target texts or different concatenation combinations (Sellam et al., 2020; Rei et al., 2020). LLM-based zero-shot approaches use prompt engineering for LLMs with the expectation of a score in the generation output (Kocmi and Federmann, 2023). In some research, attempts are made to predict evaluation scores as a weighted sum of digit tokens, where the weights are token probabilities from a Markov chain probability model (Liu et al., 2023).

The previous winner of the WMT Metrics Shared Task competition was the proprietary MetricX(Freitag et al., 2022) model, which used a regression approach. One of the state-of-the-art models in machine translation is GPT-4 with zero-shot scoring. However, due to the time consumption of its inference and the closeness of regression approaches with relatively small backbones (e.g., COMET used the base version of XLM-RoBERTa (Conneau et al., 2019) with 2.5B parameters), task-specific NLG evaluation with sophisticated tricks with vector representation and training datasets may provide better results.

Contrastive Learning Contrastive learning for NLP problems is a popular pretraining approach for improving results in downstream tasks. For example, the recent E5 model (Wang et al., 2022) is pretrained in a contrastive manner using a curated large-scale text pair dataset to solve various tasks that require a single-vector representation of texts, both after finetuning and in a zero- or few-shot manner. Another work that investigated contrastive learning for extrapolating vector representations for different tasks is InstructOR (Su et al., 2022). This model incorporates instructions in contrastive learning and achieved good results for tasks that were unseen during pretraining. The idea of knowledge transfer in the latent space may provide improvements with clean datasets and an appropriate fitting process.

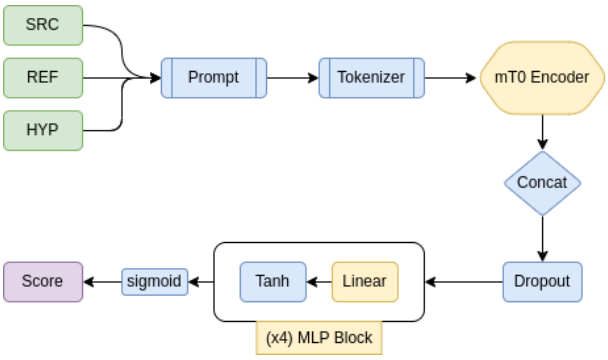


Figure 2: Architecture of the base model.

3 Method

3.1 Architecture

The main essence of our approach is to use vector representations from the encoder of the Big Science mT0 (Muennighoff et al., 2022) model as input for the Feed-Forward layer. This idea has already been proven successful in other approaches, such as COMET (Rei et al., 2020) and BLEURT (Sellam et al., 2020), and we have attempted to further improve upon it in various ways.

All our experiments use the same basic structure, as illustrated in Figure 2. First, the source text is presented as a prompt and is tokenized using the mT0 tokenizer. The resulting tokens are then processed by the mT0 encoder to obtain vector representations. Subsequently, mean pooling is applied to these vectors, and the resulting representations are passed through a Multilayer Perceptron (MLP). Prompting is necessary to present the data (source, reference, hypothesis) in a convenient format. In our approach, the prompt consists of concatenating the source, reference, and hypothesis with a separator token [sep]. It is worth noting that the mT0 tokenizer does not have a specific separator token, so another token can be selected for this purpose. Mean pooling is used to obtain sentence embeddings from the mT0 encoder representations, which are then suitable for further processing by a fully connected layer.

During the process, we tested various configurations of Multilayer Perceptron (MLP). We experi-

mented with the number of linear layers, activation functions between layers and at the end, as well as with dropouts. Insufficient linear layers resulted in a deterioration of metrics as they were unable to extract all the necessary information from the embeddings. On the other hand, adding too many layers did not have any significant impact on the results. Hence, we settled on using 4 layers. We tested Tanh and ReLU as activation functions between layers, and found that Tanh yielded slightly better results. This observation is likely due to the fact that Tanh is commonly used in neural networks for working with text embeddings. For the activation function at the end, we compared two options: sigmoid and a simple threshold approach (rounding to 100 if the result is >100 and to 0 if < 0). The use of sigmoid resulted in significantly better outcomes. In addition, we experimented with the inclusion of dropouts and found that it made sense to add only one dropout before the first linear layer after mean pooling. Otherwise, too much useful information was deleted and the MLP did not have sufficient time to extract it.

We also explored the possibility of incorporating an external «hint» in addition to the embeddings of the mT0 encoder. This approach is illustrated in scheme 1. For obtaining additional representations, we chose LaBSE (Feng et al., 2020), a state-of-the-art model in the bitext mining task, which serves as a proxy task for machine translation. In our architecture, we included an additional component responsible for preprocessing LaBSE embeddings. After obtaining these representations, they were passed through a linear layer, and the resulting vectors were concatenated with the outputs from mT0 passed through Mean Pooling. The inclusion of the linear layer after LaBSE serves as an additional degree of freedom and helps reduce the dimensionality of the vectors.

In the end, we selected a configuration that utilized a pre-trained approach combining contrastive learning mT0 and LaBSE to submit our results. This configuration demonstrated the best metrics on our test data.

3.2 Contrastive pretraining

To enhance the vector representation and address the specific characteristics of the Hebrew language, which is not as widely studied as English or German, we experimented with tuning encoder embeddings using contrastive learning. For each source text, we created two contrastive loss components: one for the reference translation and one for the machine translation. To implement this approach, we needed to specify negative examples that we wanted to be dissimilar to the source text in terms of vector representations. We used the Sentence-T5 model (Ni et al., 2021) to embed each source text and its two translations. Additionally, we constructed two ANN (Approximate Nearest Neighbor) indexes (Johnson et al., 2019): one for human references and another for machine translations. These indexes allowed us to find the K furthest points from the source texts based on the dot product. Note that for normalized vectors:

$$\|x-y\|_2^2 = 2 - 2x^T y \;\rightarrow$$
$$\min_x \left(x^T y\right) = \max_x \left((-x)^T y\right) = \min_x \left(\|(-x)-y\|_2^2\right) \tag{1}$$

The loss function is defined as the negative log likelihood with an arbitrary similarity function $\mathrm{sim}(x,y)$ (we used $\mathrm{sim}(x,y) =< x,y >$) and a temperature parameter τ. Our goal is to incorporate scaled target values of three types: SQM, DA, and MQM, with different prioritization weights in the loss function. For a given source text s, its reference translation r, and machine translation t, we have an expert degree $a^c_{s,t} \in [0,1]$ of type c with a prioritization weight γ_c. Each source text s is embedded as e_s, the reference translation r is embedded as e^+_r, and the machine translation t is embedded as e^+_t. In the case of reference translations, we denote the K furthest points from s as $\left\{e^-_{r,k}\right\}_{k=1}^K$. Similarly, in the case of machine translations, we denote the K furthest points from s as $\left\{e^-_{t,k}\right\}_{k=1}^K$.

The component for human reference:

$$\mathbf{L}\left(e_s, e^+_r, \left\{e^-_{r,k}\right\}_{k=1}^K\right) = -\log p_{s,r} \tag{2}$$

$$p_{s,r} = \frac{\exp\left(\frac{\mathrm{sim}(e_s,e^+_r)}{\tau_r}\right)}{\exp\left(\frac{\mathrm{sim}(e_s,e^+_r)}{\tau_r}\right) + \sum_{k=1}^K \exp\left(\frac{\mathrm{sim}(e_s,e^-_{r,k})}{\tau_r}\right)} \tag{3}$$

This formula is general negative log-likelihood (NLL) with temperature for self-supervised learning (Wang and Isola, 2022).

The component for machine translation:

$$\mathbf{L}\left(e_s, e_t^+, \left\{e_{t,k}^-\right\}_{k=1}^K\right) = -\log p_{s,t,c} \tag{4}$$

$$p_{s,t,c} = \frac{\exp\left(\alpha_t \frac{\text{sim}(e_s, e_t^+)}{\tau_t}\right)}{\exp\left(\alpha_t \frac{\text{sim}(e_s, e_t^+)}{\tau_t}\right) + \sum_{k=1}^K \exp\left(\frac{\text{sim}(e_s, e_{t,k}^-)}{\tau_t}\right)} \tag{5}$$

$$\alpha_t = \gamma_c a_{s,t}^c \tag{6}$$

Here we have a modified version of previous loss where we use target scores and their prior weights as temperature, but only for positive object.

Consider the derivative of the temperatured NLL loss w.r.t. to source text dot product as similarity function:

$$\frac{\left(1 - \frac{\exp\left(e_s^T e^+/\tau\right)}{Z(e_s)}\right)}{e^+\tau} - \sum_{e^-} \frac{\frac{\exp\left(e_s^T e^-/\tau\right)}{Z(e_s)}}{e^-\tau}$$

We have two separate additives with actually independent temperature coefficients. Increasing them removes the gradient changing effect and provides a pipeline for reducing the gradient step for bad and noisy translations. We can model such effect with human degrees with prioritizing ones over others.

Having a batch of quadruplets $\left\{\left(s, r, t, a_{s,t}^c\right)_n\right\}_{n=1}^N$ and using formulas above, the total loss can be written as:

$$\mathbf{L}\left(\left\{\left(s, r, t, a_{s,t}^c\right)_n\right\}_{n=1}^N\right) = \frac{1}{N}\sum_n \mathbf{L}\left(\left(s, r, t, a_{s,t}^c\right)_n\right) \tag{7}$$

$$\mathbf{L}\left(\left(s, r, t, a_{s,t}^c\right)_n\right) = \\ = \mathbf{L}\left(e_{s_n}, e_{r_n}^+, \left\{e_{r_n,k}^-\right\}_{k=1}^K\right) + \mathbf{L}\left(e_{s_n}, e_{t_n}^+, \left\{e_{t_n,k}^-\right\}_{k=1}^K\right) \tag{8}$$

Here we have an empirical risk over the batch, for each point we have two additive components for human reference and machine translation correspondingly.

3.3 Synthetic data

In this year's WMT Metrics Shared Tasks, the organizers presented us with a novel language pair: Hebrew-English. This language pair is not included in any of the available training data for MT evaluation metrics. Consequently, we believe that it was intended to test the ability of novel metrics for zero-shot transfer. To address this challenge, we made the decision to create a synthetic dataset for the Hebrew-English language pair, following the approach proposed by Rei et al. (2022b).

First, we selected a subset of English-Hebrew translations from the publicly available OPUS dataset (Tiedemann, 2012). From a total of approximately 1 million translations, we randomly chose 60,000 translations (Hebrew texts) and translated them back from Hebrew to English. To ensure a diverse range of translation quality, we selected three translation models of different sizes from the NLLB project (Costa-jussà et al., 2022): models with 600M and 1.3B parameters, which were distilled from 54B Mixture-of-Experts teacher models, as well as a 3.3B model that was trained from scratch. Each model was used to generate translations for an equally-sized portion of the dataset. Synthetic quality scores for these translations were computed as the average of scores calculated by the COMET-22 (Rei et al., 2022a) and BLEURT-20 (Sellam et al., 2020) metrics.

4 Experiments

4.1 Data

For our experiments, we utilize datasets from the previous year's WMT Metrics Shared Tasks as both training and evaluation data. These datasets provide three types of scores:

- MQM - Multidimensional Quality Metrics (Burchardt, 2013): This metric encompasses a wide range of issues that occur with translation.
- SQM - Scalar Quality Measure: This metric provides segment-level scalar ratings with document context.
- DA - Direct Assessment: This metric measures the quality of a translation on a scale from 0 to 100, based on the adequacy and fluency of the sentence.

We utilize all the available data and apply min-max scaling to rescale the score values, ensuring

they fall within the range of 0 to 1. For DA and SQM scores, we used dataset-level statistics for scaling. However, for MQM scores, we adapted the scaling to accommodate different score ranges. Specifically, the English-German and Chinese-English pairs had a range of −25 to 0, while the English-Russian pair had a range of $-\infty$ to 100.

The resulting composition of the training dataset for our experiments is as follows:

- MQM scores for WMT competitions from the years 2020 and 2021, covering 3 language pairs (en-ru, zh-en, en-de).
- SQM scores for the year 2022, covering 12 language pairs.
- DA scores for the years 2017-2022, covering 41 language pairs.

For the test set, we selected the MQM scores for the year 2022 to ensure comparability with existing metrics.

Furthermore, we included synthetic data for the Hebrew-English language pair, as described in Section 3.3. Out of the total 60,000 examples, we randomly chose 50,000 examples for the training set and the remaining 10,000 examples for the test set. Since the scores for the synthetic data were computed using existing metric models, they naturally fell within the range of 0 to 1, and no additional re-scaling was required. In total, we had 1,527,567 examples in the training set and 77,575 in the test set.

4.2 Experimental settings

All experiments were conducted with a fixed random seed. For the base of the generic model, we chose the encoder part of the mT0-large model introduced in Muennighoff et al. (2022). An MLP on top of the encoder consists of three layers with hidden sizes of 384, 96, and 1, using the hyperbolic tangent activation function. We also apply dropout with a rate of $p = 0.1$. For models that utilize embeddings, we include a resizing dense layer that projects the concatenated embeddings vector into vectors with a size of 512.

For contrastive pretraining, we once again utilize the encoder part of the mT0-large model. Contrastive examples are collected into a total batch size of 128 examples. Furthermore, we accumulate batches across four iterations, resulting in an effective batch size of 512 for each training process.

Pipeline	en-de	zh-en	en-ru	he-en
Comet-22	0.281	0.395	0.330	NA
CometKiwi	0.266	0.343	0.297	NA
Base	0.276	0.179	0.350	0.796
Base + Emb.	0.255	0.173	0.331	0.785
CL Base	0.223	0.101	0.307	0.786
CL Base + Emb.	0.222	0.105	0.315	0.792

Table 1: Experimental results on WMT22 Test Set along with our synthetic test set for He-En. **Base** model represents model that only consits of mT0-large encoder and MLP head. **CL Base** represents model that was pretrained with contrastive loss before fine-tunning.

The first two models, which are based on the original mT0-large encoder, were trained for 3 epochs with an aggressive learning rate of 2×10^{-4}. The other two models, which utilize a contrastively-pretrained encoder, were trained for 1 epoch with a learning rate of 5×10^{-5}. In both cases, the batch size was set to 8 due to the substantially larger sequence sizes.

All our experiments were conducted in a distributed data-parallel setting across 4 GPUs. The learning rate was scaled accordingly based on the number of processes.

4.3 Hardware, Computational Budget and Environmental Impact

For our experiments, we utilized the CITEC computational cluster hosted at Bielefeld University. Each node in the cluster consists of 4xA40 GPUs with 48GB of VRAM, 1xAMD EPYC 7713 64-Core CPU, and 512GB of RAM.

The total computational budget for our experiments is 175 GPU-hours (~43.75 hours per node × 4 GPUs). Considering that the A40 GPU has a power draw of 300W under full load, and the current carbon intensity of the German power grid is 510gCO2eq/kWh [3], our estimated total carbon footprint is approximately 26.775 kgCO2eq. It is important to note that this number should be considered a lower bound, as we have not accounted for the power draw of other components of the computing node, such as the CPU and cooling.

5 Results and Discussion

We trained and tested each configuration on our test data using the Kendall-τ correlation metric. The results in Table 1 show that the base configuration has the best performance in most language pairs. However, adding external semantically-informed embeddings improves the quality for the model version with the contrastive loss. We didn't manage to get better results relatively to the base model, even for the rare language pair. We think that it's due to choice of negative sampling strategy, lack of theoretical approach analysis and hyperparameter tuning. Temperature is sensitive parameter, the wrong choice of it could lead to permanent overfitting and noisy results. We need to test more natural approach with adding scaled human degrees as general temperature for all softmax components. Also we should test other approaches with metric learning, e.g. Multi-Class N-pair loss (Sohn, 2016).

6 Conclusion

This paper presents our experiments with semantically informed architectures with a regression head. This led us to conclude that the additional awareness of the encoder and extra pretraining may positively affect the model quality in these conditions. In the future, it would be possible to explore other ways to inform the model and conduct experiments with larger versions of our implemented architectures.

Limitations

While we examine a novel approach to NLG evaluation, it is important to note limitations in our research.

Firstly, due to time and computational resource constraints, we have not conducted hyperparameter search. This opens up a possibility of finding better results for reported model configurations. Additionally, we have only made one experiment with one fixed random seed for each configuration. Increasing the number of runs would improve result stability.

Acknowledgements

The NLLG group is funded by the BMBF project «Metrics4NLG» and the DFG Heisenberg grant EG 375/5-1.

[3]Data obtained from https://app.electricitymaps.com/zone/DE on September 5, 2023

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Aljoscha Burchardt. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F T. Martins. 2022. Results of wmt22 metrics shared task: Stop using bleu – neural metrics are better and more robust. *Association for Computational Linguistics*.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B Hall, Daniel Cer, and Yinfei Yang. 2021. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:2108.08877*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, José G.C. de Souza, Pedro G. Ramos, André F.T. Martins, Luisa Coheur, and Alon Lavie. 2022b. Searching for COMETINHO: The little metric that could. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 61–70, Ghent, Belgium. European Association for Machine Translation.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29.

Hongjin Su, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, Tao Yu, et al. 2022. One embedder, any task: Instruction-finetuned text embeddings. *arXiv preprint arXiv:2212.09741*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Tongzhou Wang and Phillip Isola. 2022. Understanding contrastive representation learning through alignment and uniformity on the hypersphere.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. 2020. On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1656–1671, Online. Association for Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Empowering a Metric with LLM-assisted Named Entity Annotation: HW-TSC's Submission to the WMT23 Metrics Shared Task

Zhanglin Wu*, Yilun Liu*, Min Zhang, Xiaofeng Zhao, Junhao Zhu,
Ming Zhu, Xiaosong Qiao, Jingfei Zhang, Miaomiao Ma, Yanqing Zhao
Song Peng, Shimin Tao, Hao Yang, Yanfei Jiang
Huawei Translation Service Center, Beijing, China
{wuzhanglin2,liuyilun3,zhangmin186,zhaoxiaofeng14,zhujunhao,
zhuming47,qiaoxiaosong,zhangjingfei,mamiaomiao,zhaoyanqing,
pengsong2,taoshimin,yanghao30,jiangyanfei}@huawei.com

Abstract

This paper presents the submission of Huawei Translation Service Center (HW-TSC) to the WMT23 metrics shared task, in which we submit two metrics: KG-BERTScore and HWTSC-EE-Metric. Among them, KG-BERTScore is our primary submission for the reference-free metric, which can provide both segment-level and system-level scoring. While HWTSC-EE-Metric is our primary submission for the reference-based metric, which can only provide system-level scoring. Overall, our metrics show relatively high correlations with MQM scores on the metrics tasks of previous years. Especially on system-level scoring tasks, our metrics achieve new state-of-the-art in many language pairs.

1 Introduction

Due to the expensive cost of human evaluation, automatic metrics (Freitag et al., 2022) for machine translation (MT) (Wei et al., 2021, 2022a) is critically important for MT research and development. While human evaluation is still very important, automatic metrics allow the rapid evaluation and comparison of MT systems on large collections of text and facilitate expansion to low resource languages (Li et al., 2022) and domains (Yang et al., 2021; Wu et al., 2022a). Depending on whether the references are required or not, automatic metrics are categorized into two categories: (1) reference-based metrics like BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), BERTScore (Zhang et al., 2020) and BLEURT (Sellam et al., 2020), which evaluate the hypothesis by referring to the references; (2) reference-free metrics like YiSi-2 (Lo, 2019) and COMET-QE (Rei et al., 2020, 2021), which are also referred to as quality estimation (QE). These metrics estimate the quality of hypothesis based solely on the sources, without relying on the references.

The WMT23 metrics shared task invites submissions of reference-free metrics and reference-based metrics to find automatic metric scores for translations at the segment-level and system-level. This paper presents the contribution of HW-TSC to the WMT23 metrics shared task. Slightly different from our participation last year (Liu et al., 2022a), we only submit two metrics this year. Details of our metrics (KG-BERTScore and HWTSC-EE-Metric) are illustrated in Table 1.

Metric	Category	Segment-level	System-level
KG-BERTScore	reference-free	✓	✓
HWTSC-EE-BERTScore	reference-based	✗	✓

Table 1: Details of our metrics

KG-BERTScore (Wu et al., 2022b) incorporates multilingual knowledge graph (Chen et al., 2017) into BERTScore (Zhang et al., 2019) and generates the final evaluation score by linearly combining the results of KGScore and BERTScore. Our efforts this year build on findings and observations from our participation in the WMT22 metrics shared task (Liu et al., 2022a) to further improve the accuracy of KGScore and BERTScore. The choice of a named entity (NE) annotator (Marrero et al., 2013) is critical to KGScore. With the emergence of large language models (LLMs) (Wei et al., 2022b; Kasneci et al., 2023) such as ChatGPT (Ding et al., 2022), the NE annotator seems to have one more option. Therefore, we try to use ChatGPT[1] for NE annotation and find that LLM-assisted NE annotation can empower the metric. At the same time, the selection of a QE model is crucial for BERTScore. Since COMET-QE (Rei et al., 2022) has proven to be the state-of-the-art QE model, we use it to calculate BERTScore this year.

The HWTSC-EE-Metric (Liu et al., 2022b) is developed using existing metrics with the goal of creating a more balanced scoring system at the sys-

*These authors contributed equally to this work.

[1] https://platform.openai.com

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 822–828
December 6–7, 2023.

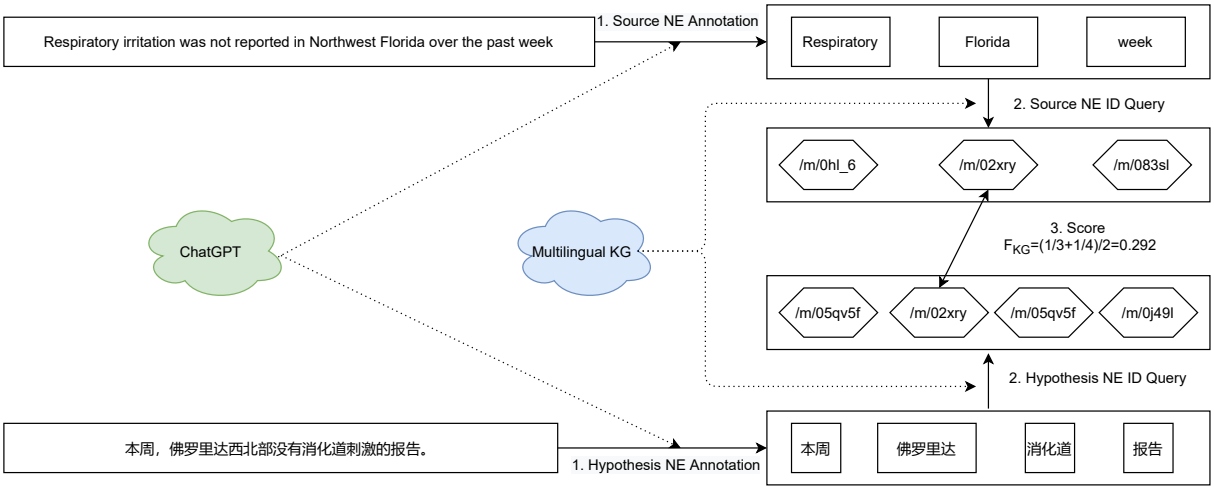


Figure 1: A Calculation Example of KGScore on English-Chinese Language Pair

tem level. This is achieved by assigning weights to segment-level scores obtained from backbone metrics. The weights are determined based on the difficulty of each segment, which is determined by the entropy of a hypothesis-reference pair. Segments with higher entropy values, indicating higher difficulty, receive larger weights in the aggregation of system-level scores by HWTSC-EE-Metric.

2 Metrics

This section introduces our metrics for WMT23 metrics shared task, including KG-BERTScore and HWTSC-EE-Metric.

2.1 KG-BERTScore

KG-BERTScore (Wu et al., 2022b) is a reference-free metric we proposed last year, which generates the final evaluation score by linearly combining the results of KGScore and BERTScore. For a given KGScore F_{KG} and BERTScore F_{BERT}, KG-BERTScore $F_{KG-BERT}$ is defined as:

$$F_{KG-BERT} = \alpha \cdot F_{KG} + (1-\alpha) \cdot F_{BERT}, \quad (1)$$

where α is an adjustable weight parameter.

We have made some improvements to the implementation details of KGScore and BERTScore, which will be described in detail below.

2.1.1 KGScore

KGScore refers to scoring based on the matching rate of NE. Figure 1 is a calculation example of KGScore on English-Chinese language pairs. The calculation process includes three steps:

Firstly, we utilize a NE Annotator to annotate NEs in the source and hypothesis sentences. Last year we used spacy[2] (Algamdi et al., 2022) as the NE annotator, but it didn't work very well. This year we try to use ChatGPT to annotate NE, and find that its effect is better than spacy, which means that LLM-assisted NE annotation is feasible.

Secondly, we match cross-lingual NE pairs by querying multilingual knowledge graphs. Google Knowledge Graph[3] (Google KG) is a general-purpose multilingual knowledge graph that we have chosen to use as always for querying NE IDs. Since same-meaning NEs in different languages share the same NE ID in Google KG, we can match cross-lingual NE pairs by NE ID. One more thing to be noted is that an NE without an ID is considered invalid and will not participate in the subsequent calculation of KGScore.

Finally, we explore using NE's matching rate to score. For a given test set with n sentence pairs, assuming that S_i is the NE numbers in the i-th source sentence, H_i is the NE numbers in the i-th hypothesis sentence, and SH_i is the number of matched cross-lingual NE pairs. The segment-level NE matching rates of the i-th source sentence and hypothesis sentence are respectively defined as:

$$F_{KGS_i} = \frac{SH_i}{S_i} \quad if\ S_i \neq 0\ else\ 1 \quad (2)$$

$$F_{KGH_i} = \frac{SH_i}{H_i} \quad if\ M_i \neq 0\ else\ 1 \quad (3)$$

[2] https://spacy.io/models

[3] https://developers.google.com/knowledge-graph

Then the segment-level calculation formula of KGScore is defined as:

$$F_{KG_i} = \frac{F_{KGS_i} + F_{KGH_i}}{2} \tag{4}$$

For the system-level KGScore, we first calculate the system-level NE matching rates of source sentences and hypothesis sentences are respectively defined as:

$$F_{KGS} = \frac{\sum_{i=1}^{n} SH_i}{\sum_{i=1}^{n} S_i} \tag{5}$$

$$F_{KGH} = \frac{\sum_{i=1}^{n} SH_i}{\sum_{i=1}^{n} H_i} \tag{6}$$

Then the system-level calculation formula of KGScore is defined as:

$$F_{KG} = \frac{F_{KGS} + F_{KGH}}{2} \tag{7}$$

2.1.2 BERTScore

BERTScore (Zhang et al., 2020) refers to scoring based on semantic similarity. We initially use Sentence-BERT (Reimers and Gurevych, 2019) to calculate the semantic similarity score between the source and hypothesis. Last year we used a reference-free HWTSC-teacher-Sim metric (Zhang et al., 2022) as BERTScore to make the score more relevant to MQM score (Lommel et al., 2014). As COMET-QE has been proven to be the state-of-the-art reference-free metric on WMT22 metrics shared task, we use the COMET-QE model[4] to score and serve as BERTScore this year.

2.2 HWTSC-EE-Metric

The HWTSC-EE-Metric, also known as the entropy-enhanced (EE) Metrics (Liu et al., 2022b), was employed in system-level shared tasks this year. Unlike traditional methods of acquiring system-level scores, EE metrics deviate from the normal approach of obtaining system-level scores via arithmetic average. EE metrics assign higher weights to difficult samples present in the evaluation set, as opposed to treating all source-reference pairs equally, as human scorers tend to do in MT evaluation. It is worth noting that simple samples can be easily translated, leading to similar human scores for different hypotheses. Conversely, challenging samples within the evaluation set play a crucial role in differentiating top candidates from inferior systems. Consequently, MT evaluation metrics should encourage systems that excel in translating difficult samples. Contrary to concerns about incorrect scoring, the use of challenging segments to evaluate MT systems has actually shown potential for improving metric performance. EE metrics, in particular, place a strong emphasis on the translation quality of difficult hypotheses and allocate higher weights to them in system-level scores.

2.2.1 Working Process of EE Metrics

EE metrics use the average qualities of hypotheses to determine the difficulty of a segment. One key measure used in this process is chunk entropy (Yu et al., 2015), which quantifies the quality of translation between the reference and the hypothesis. Higher chunk entropy indicates higher uncertainty in translation, while lower entropy suggests good confidence in the hypothesis. By calculating the entropy, easy and difficult samples can be classified accordingly through a threshold value h. In the process of aggregating scores, hypotheses are assigned weights based on their group, whether they belong to the easy or difficult category. Easy samples receive a lower weight denoted as w/N_e, while difficult samples receive a higher weight $(1-w)/N_d$. The reason for such a weight discrepancy lies in the larger number of easy hypotheses compared to difficult ones. The balance coefficient w may vary depending on the language pairs and evaluation datasets utilized. This weight assignment strategy ensures that the weights of easy samples remain significantly lower than those of difficult samples, considering the different samples in each category.

2.2.2 Enhancements to HWTSC-EE-Metric

The earlier version of EE metrics incorporates two adjustable hyperparameters, h and w, which are responsible for selecting difficult samples and assigning weights to each group, respectively. However, the presence of these hyperparameters hampers the practical application of EE metrics. Furthermore, these hyperparameters often vary across different language pairs and evaluation datasets, as evidenced by our preliminary experiment that involved up to 10 different parameters using the WMT19 evaluation set. Consequently, it becomes challenging to identify a suitable combination of hyperparameters for real-world scenarios. To address this issue, in last year's WMT metrics shared tasks (Liu et al., 2022a), we simplified the com-

[4] https://huggingface.co/Unbabel/wmt22-cometkiwi-da

putation of the system-level score by employing a normal distribution fitting approach to determine the threshold h for each translation direction. This year, we further simplified the estimation of w by using a fixed value of 0.8, as opposed to the three different configurations of w used last year. Based on the results of WMT22, we observed that the value 0.8 corresponds to an appropriate balance of weights between difficult and easy groups, as it exhibits a high correlation with human MQM scores on recent WMT test sets. Another modification in this year's HWTSC-EE-Metric is the replacement of our backbone metric from BERTScore (Zhang et al., 2019) to COMET score (Rei et al., 2022). Specifically, we adopted the model *wmt22-comet-da*[5], which is known for its robust segment-level MT evaluation capabilities, as the segment-level backbone metric for HWTSC-EE-Metric this year.

3 Experiments

This section introduces the experimental results of KG-BERTScore and HWTSC-EE-Metric on previous metrics shared tasks.

3.1 Experiment of KG-BERTScore

In order to verify the feasibility of the improved KG-BERTScore, we conduct experiments on the WMT22 metrics shared task data. Since it is time-consuming and expensive to query ChatGPT and Google Knowledge Graph API, we only verify the effect of KG-BERTScore on Chinese-English language pair. In the experiment, we first calculate F_{KGS} and F_{KGH} through NE annotation and NE pair matching, and then calculate KGScore. Next, we use COMETKiwi-22 as BERTScore to calculate the final KG-BERTScore.

We calculate the correlation of the scores of each stage (including F_{KGS}, F_{KGH}, KGscore and KG-BERTScore) with the MQM scores without considering human translation. To facilitate comparison with the official results of the WMT22 metrics shared task, the segment-level correlation adopts Kendall correlation, and the system-level correlation adopts Pearson correlation.

3.1.1 Segment-level Correlation

Table 2 shows Kendall Tau correlation of reference-free metrics with segment-level MQM scores for the WMT22 Chinese-English language pair, which is calculated without human translation. We find that KGScore has a relatively low segment-level correlation with MQM scores, while COMETKiwi-22 has a relatively high segment-level correlation with MQM scores. Therefore, when calculating KG-BERTScore, we set α to a smaller value (i.e., 0.1). Overall, the segment-level correlation between KG-BERTScore and MQM scores is only slightly higher than that of COMETKiwi-22.

Metric	Correlation
KG-BERTScore-22	0.219
COMETKiwi-22	0.364
F_{KGS}	0.017
F_{KGH}	0.055
KGScore	0.061
KG-BERTScore (α=0.1)	**0.365**

Table 2: Kendall Tau correlation of reference-free metrics with segment-level MQM scores for the WMT22 Chinese-English language pair, which is calculated without human translation.

3.1.2 System-level Correlation

Table 3 shows Pearson correlation of reference-free metrics with system-level MQM scores for the WMT22 Chinese-English language pair, which is calculated without human translation. The system-level correlation between KGScore and MQM scores is relatively close to that of COMETKiwi-22, so we set α to a larger value (i.e., 0.9). Surprisingly, the system-level correlation between KG-BERTScore and MQM scores is significantly higher than that of COMTKiwi-22.

In addition, the segment-level and system-level correlations of KGScore with MQM scores are higher than those of F_{KGS} and F_{KGH}, which indicates that both source and hypothesis NE pair matching rates should be considered when calculating KGScore.

Metric	Correlation
KG-BERTScore-22	0.743
COMETKiwi-22	0.866
F_{KGS}	0.660
F_{KGH}	0.376
KGScore	0.697
KG-BERTScore (α=0.9)	**0.947**

Table 3: Pearson correlation of reference-free metrics with system-level MQM scores for the WMT22 Chinese-English language pair, which is calculated without human translation.

[5]https://huggingface.co/Unbabel/wmt22-comet-da

Metric	En→ De (w/o Human)			Zh→ En (w/o Human)			En→ Ru (w/o Human)			En→ De (with Human)			Zh→ En (with Human)			En→ Ru (with Human)		
	r	τ	ρ	r	τ	ρ	r	τ	ρ	r	τ	ρ	r	τ	ρ	r	τ	ρ
WMT21-news																		
BERTScore	0.911	0.795	**0.945**	0.577	0.308	0.484	0.776	0.538	0.692	0.181	0.441	0.500	0.382	0.295	0.439	0.540	0.417	0.485
COMET	0.812	0.590	0.819	0.545	0.359	0.401	0.774	0.538	0.688	0.349	0.559	0.804	0.425	0.333	0.386	0.751	0.617	0.782
EE-BERTScore-0.3	0.874	**0.846**	**0.945**	**0.637**	**0.487**	**0.626**	0.621	0.451	0.622	0.182	0.485	0.512	0.384	**0.410**	**0.521**	0.569	0.317	0.435
EE-BERTScore-0.5	0.898	**0.846**	**0.945**	0.595	0.359	0.511	0.717	0.495	0.701	0.183	0.500	0.517	0.382	0.352	0.457	0.562	0.383	0.491
EE-BERTScore-0.8	**0.919**	0.769	0.923	0.526	0.256	0.462	0.809	**0.604**	**0.754**	0.184	0.456	0.532	0.380	0.276	0.429	0.548	0.467	0.526
HWTSC-EE-Metric	0.816	0.615	0.819	0.474	0.359	0.462	**0.814**	0.582	0.727	**0.380**	**0.574**	**0.806**	**0.427**	0.333	0.454	**0.761**	**0.683**	**0.821**
WMT21-tedtalks																		
BERTScore	0.465	0.256	0.319	0.634	0.055	0.134	0.826	0.626	0.793	0.541	0.363	0.455	-0.634	-0.086	-0.079	0.659	0.676	0.832
COMET	0.764	0.436	0.604	0.620	**0.143**	0.196	**0.878**	0.692	**0.868**	0.626	0.516	0.684	-0.638	**-0.010**	-0.029	0.784	**0.733**	**0.893**
EE-BERTScore-0.3	0.560	0.333	0.473	0.321	0.055	0.125	0.687	0.451	0.626	0.553	0.429	0.578	-0.775	-0.086	-0.086	-0.568	0.219	0.289
EE-BERTScore-0.5	0.558	0.333	0.445	0.534	0.077	0.143	0.750	0.495	0.679	0.549	0.429	0.556	-0.719	-0.067	-0.071	-0.538	0.276	0.361
EE-BERTScore-0.8	0.495	0.359	0.478	**0.645**	0.077	0.134	0.829	0.692	0.829	0.543	0.451	0.582	-0.617	-0.067	-0.079	**0.805**	0.714	0.857
HWTSC-EE-Metric	**0.799**	**0.538**	**0.742**	0.633	**0.143**	**0.213**	0.869	**0.851**	0.692	**0.653**	**0.604**	**0.793**	**-0.593**	**-0.010**	**-0.014**	-0.005	0.467	0.504

Table 4: Correlations with system-level human MQM scores on datasets of WMT21 news and WMT21 tedtalks. EE-BERTScore-$*$ represents our last year's submission in WMT22. HWTSC-EE-Metric represents our submission in WMT23. **With Human** indicates evaluation on MT systems and human translations, and **w/o Human** indicates MT systems only. Best correlations are marked in bold.

3.1.3 Effect of Different Weights

KG-BERTScore generates the final evaluation score by linearly combining the results of KGScore and BERTScore. α is an adjustable weight parameter in the linear combination formula, which affects the correlation between KG-BERTScore and MQM scores. To analyze the effect of α value, we calculate the segment-level and system-level correlations of KG-BERTScore and MQM scores under different α values for the WMT22 Chinese-English language pair. The result is shown in Figure 2. The segment-level correlation between KG-BERTScore and MQM scores is highest when the α value is 0.1, and the system-level correlation between KG-BERTScore and MQM scores is the highest when the α value is 0.9. That is to say, when the correlation between KGScore and MQM scores is relatively low, α should take a smaller value, otherwise, α should set a larger value.

On the WMT23 metrics shared task, we cannot know the MQM score in advance. Therefore, we refer to the above experimental settings, and set α to 0.1 and 0.9 on the segment-level and system-level metrics shared tasks, respectively. In addition, we do not calculate KG-Score and set α to 0 on non-MQM language pairs due to the slow speed of accessing ChatGPT.

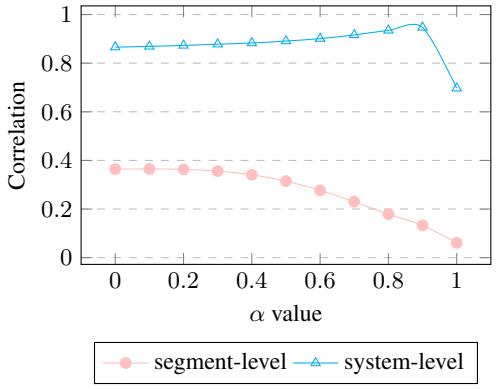


Figure 2: The segment-level and system-level correlations between KG-BERTScore and MQM scores under different α values for the WMT22 Chinese-English.

3.2 Experiment of HWTSC-EE-Metric

To evaluate the performance of the HWTSC-EE-Metric, a series of experiments were conducted primarily on the WMT21 test sets using the MQM scores as the human scoring standard. To investigate the impact of using human translations as part of the system, the results obtained from two sets of systems for each language pair are compared. The evaluation was based on three coefficients: Pearson's correlation coefficient (r), Kendall's τ, and Spearman's ρ, which are used to assess the system-level correlations with human evaluations.

Table 4 presents a performance comparison between the HWTSC-EE-Metric (our submission in WMT23), EE-BERTScore (our submission in WMT22), and two standard metrics (BERTScore and COMET). The HWTSC-EE-Metric demonstrates higher overall correlations with human MQM evaluations compared to its backbone, the standard COMET score. Furthermore, out of the 36 comparison terms, the HWTSC-EE-Metric achieves the best performance in 20 cases. This strong performance indicates the effectiveness of our entropy-based enhancing strategy and parameter estimation approach.

As EE metrics evaluate a system based not only on the individual system itself but also on other participating systems, the inclusion of human trans-

lations may influence the performance of EE metrics. As shown in Table 4, most metrics exhibit a decline in performance when human translations are included. The improvements of the HWTSC-EE-Metric in correlations with MQM are not consistently steady, which aligns with the findings of (Freitag et al., 2021) that most metrics struggle to accurately score translations that differ from MT systems. However, we observed that the HWTSC-EE-Metric mitigates the performance reduction of COMET in some cases (*e.g.*, En→ De in WMT21 datasets), but there are also instances where the HWTSC-EE-Metric does not improve COMET in terms of correlations (*e.g.*, En→ Ru in WMT21 TED talks). Overall, when human translations are included as additional outputs, EE metrics tend to be less robust and provide a less significant improvement over standard metrics.

4 Conclusion

This paper presents HW-TSC's submission to the WMT23 metrics shared task, in which we summit a reference-free metric (KG-BERTScore) and a reference-based metric (HWTSC-EE-Metric). We have made some improvements to these two metrics compared to last year's submission. One of the most critical improvements is on KG-BERTScore, we empower the metric with LLM-assisted NE annotations, significantly improving its correlation with MQM scores. The experimental results on previous WMT metrics tasks show great effectiveness of our research direction and the superiority of our metrics.

References

Shabbab Algamdi, Abdullah Albanyan, Sayed Khushal Shah, and Zeenat Tariq. 2022. Twitter accounts suggestion: Pipeline technique spacy entity recognition. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 5121–5125. IEEE.

Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo. 2017. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 1511–1517.

Bosheng Ding, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq R. Joty, and Boyang Li. 2022. Is gpt-3 a good data annotator? In *Annual Meeting of the Association for Computational Linguistics*.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André FT Martins. 2022. Results of wmt22 metrics shared task: Stop using bleu–neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the wmt21 metrics shared task: Evaluating metrics with expert-based human evaluations on ted and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Shaojun Li, Yuanchang Luo, Daimeng Wei, Zongyao Li, Hengchao Shang, Xiaoyu Chen, Zhanglin Wu, Jinlong Yang, Zhiqiang Rao, Zhengzhe Yu, et al. 2022. Hw-tsc systems for wmt22 very low resource supervised mt task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1098–1103.

Yilun Liu, Xiaosong Qiao, Zhanglin Wu, Chang Su, Min Zhang, Yanqing Zhao, Song Peng, Shimin Tao, Hao Yang, Ying Qin, et al. 2022a. Partial could be better than whole. hw-tsc 2022 submission for the metrics shared task. *WMT 2022*, page 549.

Yilun Liu, Shimin Tao, Chang Su, Min Zhang, Yanqing Zhao, and Hao Yang. 2022b. Part represents whole: Improving the evaluation of machine translation system using entropy enhanced metrics. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 296–307.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, pages 0455–463.

Mónica Marrero, Julián Urbano, Sonia Sánchez-Cuadrado, Jorge Morato, and Juan Miguel Gómez-Berbís. 2013. Named entity recognition: fallacies,

challenges and opportunities. *Computer Standards & Interfaces*, 35(5):482–489.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Daimeng Wei, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Xiaoyu Chen, Hengchao Shang, Jiaxin Guo, Minghan Wang, Lizhi Lei, Min Zhang, et al. 2021. Hw-tsc's participation in the wmt 2021 news translation shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 225–231.

Daimeng Wei, Zhiqiang Rao, Zhanglin Wu, Shaojun Li, Yuanchang Luo, Yuhao Xie, Xiaoyu Chen, Hengchao Shang, Zongyao Li, Zhengzhe Yu, et al. 2022a. Hw-tsc's submissions to the wmt 2022 general machine translation shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 403–410.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022b. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Zhanglin Wu, Jinlong Yang, Zhiqiang Rao, Zhengzhe Yu, Daimeng Wei, Xiaoyu Chen, Zongyao Li, Hengchao Shang, Shaojun Li, Ming Zhu, et al. 2022a. Hw-tsc translation systems for the wmt22 biomedical translation task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 936–942.

Zhanglin Wu, Min Zhang, Ming Zhu, Yinglu Li, Ting Zhu, Hao Yang, Song Peng, and Ying Qin. 2022b. KG-BERTScore: Incorporating Knowledge Graph into BERTScore for Reference-Free Machine Translation Evaluation. In *11th International Joint Conference on Knowledge Graphs, IJCKG2022*. To be publiushed.

Hao Yang, Zhanglin Wu, Zhengzhe Yu, Xiaoyu Chen, Daimeng Wei, Zongyao Li, Hengchao Shang, Minghan Wang, Jiaxin Guo, Lizhi Lei, et al. 2021. Hw-tsc's submissions to the wmt21 biomedical translation task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 879–884.

Hui Yu, Xiaofeng Wu, Wenbin Jiang, Qun Liu, and Shouxun Lin. 2015. Improve the evaluation of translation fluency by using entropy of matched subsegments. *CoRR*, abs/1508.02225.

Min Zhang, Hao Yang, Shimin Tao, Yanqing Zhao, Xiaosong Qiao, Yinlu Li, Chang Su, Minghan Wang, Jiaxin Guo, Yilun Liu, and Ying Qin. 2022. Incorporating multilingual knowledge distillation into machine translation evaluation. In *The 16th China Conference on Knowledge Graph and Semantic Computing, CCKS2022*. To be publiushed.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Unify word-level and span-level tasks: NJUNLP's Participation for the WMT2023 Quality Estimation Shared Task

Xiang Geng[1], **Zhejian Lai**[1], **Yu Zhang**[1], **Shimin Tao**[2], **Hao Yang**[2], **Jiajun Chen**[1], **Shujian Huang**[1]*
[1] National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[2] Huawei Translation Services Center, Beijing, China
{gx, laizj, zhangy}@smail.nju.edu.cn, {taoshimin, yanghao30}@huawei.com
{chenjj, huangsj}@nju.edu.cn

Abstract

We introduce the submissions of the NJUNLP team to the WMT 2023 Quality Estimation (QE) shared task. Our team submitted predictions for the English-German language pair on all two sub-tasks: (i) sentence- and word-level quality prediction; and (ii) fine-grained error span detection. This year, we further explore pseudo data methods for QE based on NJUQE framework[1]. We generate pseudo MQM data using parallel data from the WMT translation task. We pre-train the XLMR large model on pseudo QE data, then fine-tune it on real QE data. At both stages, we jointly learn sentence-level scores and word-level tags. Empirically, we conduct experiments to find the key hyper-parameters that improve the performance. Technically, we propose a simple method that covert the word-level outputs to fine-grained error span results. Overall, our models achieved the best results in English-German for both word-level and fine-grained error span detection sub-tasks by a considerable margin.

1 Introduction

Quality Estimation (QE) of Machine Translation (MT) is a task to estimate the quality of translations at run-time without access to reference translations (Specia et al., 2018). There are two sub-tasks in WMT 2023 QE shared task[2]: (i) sentence- and word-level quality prediction; and (ii) fine-grained error span detection. We participated in all two sub-tasks for the English-German (EN-DE) language pair. The annotation of EN-DE is multi-dimensional quality metrics (MQM) [3], aligned with the WMT 2023 Metrics shared task. The MQM annotation provides error spans with fine-grained categories and severities by human translators.

Inspired by DirectQE (Cui et al., 2021) and CLQE (Geng et al., 2023), we further explore pseudo data methods for QE based on the NJUQE framework. We generate pseudo MQM data using parallel data from the WMT translation task. Specifically, we replace the reference tokens with these tokens sampled from translation models. To simulate translation errors with different severities, we sample tokens with lower generation probabilities for worse errors (Geng et al., 2022). We pre-train the XLMR (Conneau et al., 2020) large model on pseudo MQM data, then fine-tune it on real QE data. At both stages, we jointly learn sentence-level scores (MSE loss and margin ranking loss) and word-level tags (cross-entropy loss).

For task (i), the QE model outputs the sentence scores and the "OK" probability of each token. For task (ii), we set different thresholds for the "OK" probability to predict fine-grained severities. We regard consecutive "BAD" tokens as a whole span and take the worse severity of each token as the result. We train different models with different parallel data and ensemble their results as the final submission.

Overall, we summarize our contribution as follows:

- Empirically, we conduct experiments to find the key hyper-parameters that improve the performance.
- Technically, we propose a simple method that converts the word-level outputs to fine-grained error span results.

Our system obtains the best results in English-German for both word-level and fine-grained error span detection sub-tasks with an MCC of 29.7 (+4.1 than the second best system) and F1 score of 28.4 (+1.1) respectively. We rank 2nd place on sentence-level sub-tasks with a Spearman score of 47.9 (-0.4 than the best system).

* Corresponding Author.
[1] https://github.com/NJUNLP/njuqe
[2] https://wmt-qe-task.github.io
[3] https://themqm.org

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 829–834
December 6–7, 2023.

Source	Government Retires 15 More Senior Tax Officials On Graft Charges	
Translation	Regierung zieht 15 weitere leitende Steuerbeamte wegen Graft-Vorwürfen zurück	
Translation Back	Government withdraws 15 more senior tax officials over graft allegations	
Tags	OK BAD OK OK OK OK OK BAD OK	
MQM Score	0.3333	
Annotation ID	**Character-level Indices of Error Span**	**Severity**
Span 1	10:15	Major
Span 2	55:70	Minor

Table 1: An example from the WMT2023 English-German MQM dataset. We mark the error span with red color. The translation back is generated by Google Translate.

2 Background

Given a source language sentence X and a target language translation $\hat{Y} = \{y_1, y_2, \ldots, y_n\}$ with n tokens, the MQM annotation provides error spans with fine-grained categories and severities (minor, major, and critical) by human translators. The MQM score sums penalties for each error severity and then normalizes the result by translation length:

$$\text{MQM} = 1 - \frac{n_{\text{minor}} + 5n_{\text{major}} + 10n_{\text{critical}}}{n}, \quad (1)$$

where n_{severity} denotes the number of each error severity and n denotes the translation length.

As shown in table 1, participating systems are required to predict tags $G = \{g_1, g_2, \ldots, g_n\}$ of each word and MQM score m for sub-task (i), where the binary label $g_j \in \{\text{OK}, \text{BAD}\}$ is the quality label for the word translation y_j. For sub-task (ii), we need to predict both the character-level start and end indices of every error span as well as the corresponding error severity. The primary metrics of sentence-level, word-level, and span detection subtasks are Spearman's rank correlation coefficient, Matthews correlation coefficient (MCC)[4], and F1-score respectively[5].

3 Methodology

Generally, we unite the sub-tasks (i) and (ii) as follows:

- We generate pseudo MQM data for sub-task (i) using parallel data and translation models as shown in the left of figure 1.
- We pre-train the QE model with pseudo data and fine-tune it with real QE data for sub-task (i) as shown in the right of figure 1.
- We ensemble the results of models trained with different parallel data for sub-task (i).
- We convert word-level probabilities for sub-task (i) to error span and fine-grained severities for sub-task (ii).

3.1 Pseudo MQM Data

We adopt the pseudo MQM data method described in (Geng et al., 2022).

3.1.1 Corrupting

Given a parallel pair (X, Y), we corrupt the reference Y as shown in figure 2:

- We sample the number of spans t according to the distribution of WMT2022 QE EN-DE valid set (Zerva et al., 2022a).
- According to the distribution of WMT2022 QE EN-DE valid set, we sample the length of each span n_i one by one to ensure that the total length is less than reference length n.
- We randomly sample the start indices for i-th span in $[\text{EOL}_i, n - \sum_{j=i}^{t} n_j]$ to ensure each span lie in the sentence, where EOL_i is the end indices of last span ($\text{EOL}_0 = 0$).
- We sample the severity of each span according to the distribution of a WMT2022 QE EN-DE valid set.
- We randomly insert or remove some tokens in each span to simulate over- and under-translations.

[4] https://github.com/sheffieldnlp/qe-eval-scripts/tree/master

[5] https://github.com/WMT-QE-Task/wmt-qe-2023-data/blob/main/task_2/evaluation

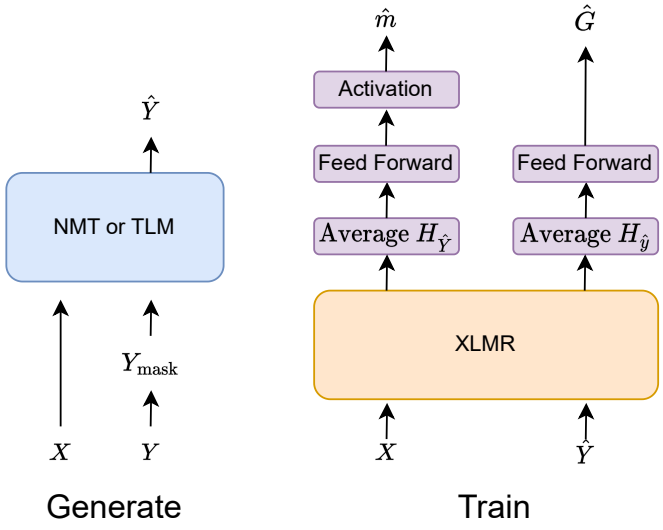


Figure 1: Illustration of the whole procedure.

- We tag tokens on the right of the omission errors and tokens that are not aligned with reference tokens as "BAD". The rest tokens are tagged as "OK". We calculate the MQM score using Eq. 1 based on the sampled severities.

3.1.2 Fixing

To generate pseudo translations, we replaced these error tokens with the "mask" symbol and sampled these tokens with neural machine translation (NMT) model (Vaswani et al., 2017) or translation language model (TLM) (Conneau and Lample, 2019). For the NMT model, we generate these error tokens from left to right with teacher forcing, while the TLM model generates these tokens parallel. To simulate errors of different severities, we sample tokens with lower generation probabilities for graver pseudo errors. To generate diverse pseudo translations, we random sample one of the tokens with the top k generation probability as the error token. In practical, we use $k = 2, 10, 100$ for minor, major, and critical errors, respectively.

3.2 Pre-training and Fine-tuning

3.2.1 QE Model

Since the pre-train models significantly improve MT evaluation performance (Rei et al., 2022; Zerva et al., 2022b), we use the XLMR large model (f) as the model backbone. To obtain the features conditioned on source sentences, we input the concatenation of source sentences and translations:

$$H_X, H_{\hat{Y}} = f(X, \hat{Y}). \tag{2}$$

Then, we average the representations $H_{\hat{Y}}$ of all target tokens as the sentence score representation H_{sent}.

$$H_{\text{sent}} = \text{Average}(H_{\hat{Y}}) \tag{3}$$

The sentence score representation passes through one linear layer and an optional activation function σ to output the score prediction $\hat{m}$.

$$\hat{m} = \sigma(\text{FFN}(H_{\text{sent}})), \tag{4}$$

where we set σ as the Sigmoid function or null. We average sub-tokens' representations as the representation of the whole word. We input the word representations H_{word} to one linear layer and softmax function to predict binary labels:

$$\hat{G} = \text{softmax}(\text{FFN}(H_{\text{word}})). \tag{5}$$

3.2.2 QE Loss

Following the multi-task learning framework for QE (Zerva et al., 2021), we joint learn the sentence- and word-level tasks. We use two loss functions for the sentence-level task: the margin ranking loss and the mean square error (MSE) loss. The margin ranking loss is defined as follows:

$$L_{\text{Rank}} = \max(0, -r(\hat{m}^i - \hat{m}^j) + \epsilon), \tag{6}$$

where $\hat{m}^i$ and $\hat{m}^j$ denote the output scores of i-th and j-th translations from current batch; r denotes the rank label, $r = 1$ if $m^i > m^j$, $r = -1$ if $m^i < m^j$; ϵ denotes the margin, we set $\epsilon = 0.03$ for all experiments. As shown in (Geng et al.,

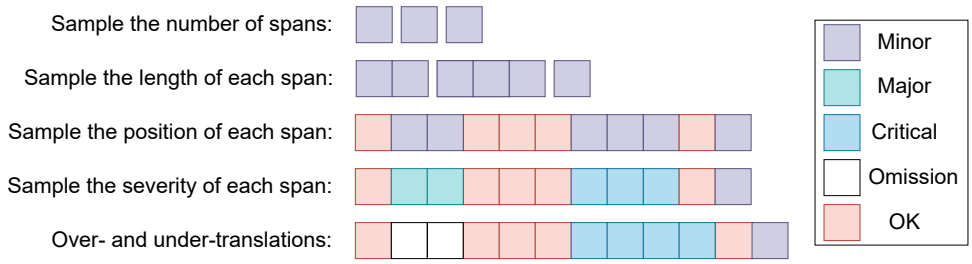


Figure 2: Illustration of the pseudo MQM data method (Geng et al., 2022). The word-level tags of this pseudo translation are annotated as "OK BAD OK OK BAD BAD BAD BAD OK BAD" and the MQM score is -0.6.

2022), the ranking loss is critical to achieving good performance. And the MSE loss is defined as:

$$L_{\text{MSE}} = \text{MSE}(m, \hat{m}). \tag{7}$$

We use cross-entropy (CE) loss for the word-level task:

$$L_{\text{CE}} = \sum_{i=1}^{n} \text{CE}(g_i, \hat{g}_i), \tag{8}$$

where $\hat{g}_i$ denotes the tag predicted for i-th word. The final QE loss function is the weighted sum of previous loss functions:

$$L_{\text{QE}} = L_{\text{CE}} + \alpha L_{\text{MSE}} + \beta L_{\text{Rank}}, \tag{9}$$

where α and β denote the weights for different loss functions. We use the Eq. 9 for both pre-training and fine-tuning.

3.3 Ensemble

We generate one pseudo MQM data for each parallel pair. We train different QE models with different pseudo MQM data and ensemble their results as the final submission. For the sentence-level task, we calculate the z-scores of each output and the average of these z-scores as the predictions. For the word-level task, we use QE models to output "OK" probabilities $P = \{p_1, p_2, \ldots, p_n\}$, where p_i denotes the "OK" probability for i-the word in the translation. Then, we average "OK" probabilities and set a threshold ϵ_{BAD} to decide whether the word is "BAD":

$$\hat{g}_i = \begin{cases} \text{OK} & \text{if } p_i > \epsilon_{\text{BAD}} \\ \text{BAD} & \text{if } p_i \leq \epsilon_{\text{BAD}} \end{cases} \tag{10}$$

3.4 Sub-task (ii)

To unite the word-level sub-task and fine-grained error span detection sub-task, we propose a simple method that covert the word-level outputs to fine-grained error span results. Based on the ensemble "OK" probabilities, we set two thresholds ϵ_{major} and ϵ_{minor}. Then, we can output the fine-grained error tags $S = \{s_1, s_2, \ldots, s_n\}$, where p_i as follows:

$$\hat{s}_i = \begin{cases} \text{OK} & \text{if } p_i > \epsilon_{\text{minor}} \\ \text{Minor} & \text{if } \epsilon_{\text{Major}} < p_i \leq \epsilon_{\text{Minor}} \\ \text{Major} & \text{if } p_i \leq \epsilon_{\text{Major}} \end{cases} \tag{11}$$

Finally, we regard consecutive error tokens as a whole span and take the worst severity of error tokens as the span severity. As recommended by the reviewer, we also try to take the majority category as the span severity. However, we found that only one prediction changed from"major" to "minor". That may be because the task is imbalanced and there are more "major" errors. As a result, this strategy achieves the same F1-score as the previous one.

4 Experiments

4.1 Implementation Details

We use parallel data from the WMT translation task to generate the pseudo MQM data. We use the WMT2022 QE EN-DE dataset and the WMT2022 Metric EN-DE dataset for fine-tuning. We also incorporate the post-editing annotation EN-DE datasets (WMT17, 19, and 20) to warm up the QE model.

We implement our system based on the NJUQE framework, which is built on the Fairseq(-py) (Ott et al., 2019) toolkit. We use NVIDIA V100 GPUs to conduct our experiments. To search the hyper-parameters, we utilize the grid search method. All experiments set the random seed as 1. We set $\alpha = 1$ and $\beta = 1000$ for both pre-training and fine-tuning. When pre-training, we use four GPUs. We set the learning rate to 1e-5, the maximum number of

σ	Spearman
w/o σ	50.02
sigmoid	52.41

Table 2: Results on the validation set of WMT2022 QE EN-DE task with different normalize function σ.

tokens in a batch to 1400 and update the parameters every four batches. We evaluate the model every 600 updates and perform early stopping if the validation performance does not improve for the last ten runs. When fine-tuning, we use one GPU. we set the learning rate to 1e-6, the maximum number of sentences in a batch to 20. We evaluate the model every 300 updates and perform early stopping if the validation performance does not improve for the last ten runs.

4.2 Results

We achieve the best results on EN-DE for both word-level and fine-grained error span detection sub-tasks with an MCC of 29.7 (+4.1 than the second best system) and F1 score of 28.4 (+1.1) respectively. We rank 2nd place on sentence-level sub-tasks with a Spearman score of 47.9 (-0.4 than the best system).

5 Analysis

In this section, we show some key hyper-parameters that improve the performance.

5.1 The normalize function σ

Although the MSE loss improves sentence-level performance, we need to avoid the over-fitting of score predictions. We set the normalize function σ as the sigmoid function to provide smooth gradients. As shown in table 2, we achieve better sentence-level performance by using the sigmoid function.

5.2 Dropout Rate of the Output Layers

We also use the dropout method (Gal and Ghahramani, 2016) on the output layers to avoid over-fitting. Table 3 shows that the QE model obtains better performance when we set the dropout rate as 0.2.

Dropout Rate	Spearman
0	52.41
0.1	52.93
0.2	53.11
0.3	52.15

Table 3: Results on the validation set of WMT2022 QE EN-DE task with different dropout rate.

6 Conclusion

We present NJUNLP's work to the WMT 2023 Shared Task on Quality Estimation. In this work, we generate pseudo MQM data using parallel data. We pre-train the XLMR large model on pseudo MQM data, then fine-tune it on real QE data. At both stages, we jointly learn sentence-level scores and word-level tags. Empirically, we conduct experiments to find the key hyper-parameters that improve the performance. Technically, we propose a simple method that covert the word-level outputs to fine-grained error span results. Overall, our models achieved the best results in English-German for both word-level and fine-grained error span detection sub-tasks by a considerable margin.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments. Shujian Huang is the corresponding author. This work is supported by National Science Foundation of China (No. 62376116, 62176120), the Liaoning Provincial Research Foundation for Basic Research (No. 2022-KF-26-02).

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Qu Cui, Shujian Huang, Jiahuan Li, Xiang Geng, Zaixiang Zheng, Guoping Huang, and Jiajun Chen. 2021. Directqe: Direct pretraining for machine translation quality estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12719–12727.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Xiang Geng, Yu Zhang, Shujian Huang, Shimin Tao, Hao Yang, and Jiajun Chen. 2022. Njunlp's participation for the wmt2022 quality estimation shared task. *WMT 2022*, page 615.

Xiang Geng, Yu Zhang, Jiahuan Li, Shujian Huang, Hao Yang, Shimin Tao, Yimeng Chen, Ning Xie, and Jiajun Chen. 2023. Denoising pre-training for machine translation quality estimation with curriculum learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12827–12835.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Ricardo Rei, Marcos Treviso, Nuno M Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José GC de Souza, Taisiya Glushkova, Duarte M Alves, Alon Lavie, et al. 2022. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. *WMT 2022*, page 634.

Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. 2018. Quality estimation for machine translation. *Synthesis Lectures on Human Language Technologies*, 11(1):1–162.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems 30*.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022a. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Chrysoula Zerva, Taisiya Glushkova, Ricardo Rei, and André F. T. Martins. 2022b. Disentangling uncertainty in machine translation evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8622–8641, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chrysoula Zerva, Daan van Stigt, Ricardo Rei, Ana C Farinha, Pedro Ramos, José G. C. de Souza, Taisiya Glushkova, Miguel Vera, Fabio Kepler, and André F. T. Martins. 2021. IST-unbabel 2021 submission for the quality estimation shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 961–972, Online. Association for Computational Linguistics.

HW-TSC 2023 Submission for the Quality Estimation Shared Task

Yuang Li, Chang Su, Ming Zhu, Mengyao Piao, Xinglin Lyu, Min Zhang, Hao Yang
Huawei Translation Services Center, China
{liyuang3, suchang8, zhuming47, piaomengyao1,
lvxinglin, zhangmin186, yanghao30}@huawei.com

Abstract

Quality estimation (QE) is an essential technique to assess machine translation quality without reference translations. In this paper, we focus on Huawei Translation Services Center's (HW-TSC's) submission to the sentence-level QE shared task, named Ensemble-CrossQE. Our system uses CrossQE, the same model architecture as our last year's submission, which consists of a multilingual base model and a task-specific downstream layer. The input is the concatenation of the source and the translated sentences. To enhance the performance, we finetuned and ensembled multiple base models such as XLM-R, InfoXLM, RemBERT and CometKiwi. Moreover, we introduce a new corruption-based data augmentation method, which generates deletion, substitution and insertion errors in the original translation and uses a reference-based QE model to obtain pseudo scores. Results show that our system achieves impressive performance on sentence-level QE test sets and ranked the first place for three language pairs: English-Hindi, English-Tamil and English-Telegu [1]. In addition, we participated in the error span detection task. The submitted model outperforms the baseline on Chinese-English and Hebrew-English language pairs.

1 Introduction

Quality estimation (QE) involves automatically scoring machine translation outputs without depending on reference translations (Specia et al., 2018). In the WMT 2023 QE shared task, there are two subtasks — quality estimation and fine-grained error span detection and each task involves several language pairs. Our team — Huawei Translation Services Center (HW-TSC) — participated in the sentence-level quality prediction and the fine-grained error span detection tasks over all language pairs except the zero-shot language pair. Fine-tuning pre-trained language models, which offers abundant semantic information, has become the paradigm for QE tasks (Rei et al., 2020). In this paper, we describe Ensemble-CrossQE, HW-TSC's system for sentence-level QE task, which leverages multiple pre-trained language models and data augmentation technique. Our system designs can be summarized as follow:

- **Model**: We employed our previous year's architecture, CrossQE (Tao et al., 2022), as the foundation. For every language pair, models were individually fine-tuned. Additionally, we used CometKiwi (Rei et al., 2022), a multilingual QE model and fine-tuned it for single language pairs.
- **Data augmentation**: The original training dataset was augmented with a novel corruption-based approach. A reference-based QE model was used to generate pseudo scores for corrupted translations by taking the original translation as reference and a corrupted translation as the new translation.
- **Ensemble**: For each language pair, 12 checkpoints were considered for the final ensemble. These checkpoints originated from four base models: XLM-R (Conneau et al., 2020), InfoXLM (Chi et al., 2021), RemBERT (Chung et al., 2020), and CometKiwi (Rei et al., 2022), and three training dataset configurations: original dataset, augmented dataset, and augmented dataset followed by the original dataset. The ensemble weight for each checkpoint was optimized with Optuna (Akiba et al., 2019). On average, eight checkpoints were used per language pair after optimization.

Our system achieves remarkable results and outperforms the baseline given by the competition organizer by a large margin. Additionally, we provide

[1] http://www2.statmt.org/wmt23/quality-estimation-task_results.html

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 835–840
December 6–7, 2023.

detailed results of each model with and without data augmentation in Table 1. To analyze the importance of each model in the ensemble, we present the ensemble weights in Figure 1 and 2. It is worth noting that the models fine-tuned with the proposed data augmentation technique were assigned higher weights in the ensemble.

2 Background

2.1 Task Description [2]

Sentence-level QE with direct assessment (DA) anotations: The goal is to predict the quality score for each source-target sentence pair. The golden-truth quality scores were obtained from human translators who rated each translation from 0 to 100. The scores from three or four translators were normalized and averaged to get the final score. This year's QE shared task has five language pairs with DA quality scores: English-Marathi (en-mr), English-Hindi (en-hi), English-Tamil (en-ta), English-Telegu (en-te) and English-Gujarati (en-gu). Only en-mr has 26,000 training samples, while the other languages have just 7,000 training samples each.

Sentence-level QE with multi-dimensional quality metrics (MQM) anotations: The goal is to predict the quality score for each source-target sentence pair. MQM can be used to identify quality issues in translation products, classify them against a shared, open and standardized error typology, and generate quality measures that can be used to gauge how well the translation product meets quality requirements. Calculating different scores by error type, the summing penalties for each MQM error category are +1 point for minor errors +5 points for major errors, and +10 points for critical errors. This year's QE shared task has two language pairs with MQM quality scores: English-German(en-de) and Chinese-English(zh-en). The en-de has 28900 training samples and zh-en has 35300 training samples.

Fine-grained error span detection: Participants of this task need to identify the error span (start and end indices) and the error severity (major or minor).

2.2 Base Models

- **XLM-R** (Conneau et al., 2020): A transformer-based masked language model trained on a massive multilingual corpus with more than two terabytes of data.
- **InfoXLM** (Chi et al., 2021): A cross-lingual pre-trained model that leverages multilingual masked language modeling, translation language modeling and cross-lingual contrast learning.
- **RemBERT** (Chung et al., 2020): A rebalanced mBERT model with factorization of the embedding layers. The input embeddings are smaller and kept for fine-tuning, while the output embeddings are larger and discarded after pre-training.
- **CometKiwi** (Rei et al., 2022): A multilingual reference-free QE model that uses a regression approach and is built on top of InfoXLM. It has been trained on direct assessments from WMT17 to WMT20 and the MLQE-PE corpus.

3 Method

3.1 Model Architecture

3.1.1 Task1: Sentence-level QE with direct assessment (DA) and multi-dimensional quality metrics (MQM) anotations

As shown in Equation 1 and 2, the embeddings of source sentence $\mathbf{s}$ and translated sentence $\mathbf{t}$ are concatenated in both orders $[\mathbf{s}, \mathbf{t}]$ and $[\mathbf{t}, \mathbf{s}]$ to form the input of pre-trained model f_{base}. The output token-level embedding sequences are processed by an average pooling layer to obtain vector reprsentations $\mathbf{h}_{s1}$ and $\mathbf{h}_{t1}$ for source and translation respectively. These feature vectors are enhanced by taking their absolute difference and element-wise multiplication, as shown in Equation 3 and 4. Finally, all feature vectors are concatenated and fed into a regression head that predicts the final score y (Equation 5). This architecture enables information exchange between source and translated sentences at an early stage of the network and has proven to be significantly more effective than combining cross-lingual information after the pre-trained model.

$$\mathbf{h}_{s1}, \mathbf{h}_{t1} = f_{base}([\mathbf{s}, \mathbf{t}]) \quad (1)$$

$$\mathbf{h}_{t2}, \mathbf{h}_{s2} = f_{base}([\mathbf{t}, \mathbf{s}]) \quad (2)$$

$$\mathbf{f}_1 = [\mathbf{h}_{s1}, \mathbf{h}_{t1}, |\mathbf{h}_{s1} - \mathbf{h}_{t1}|, \mathbf{h}_{s1} \odot \mathbf{h}_{t1}] \quad (3)$$

$$\mathbf{f}_2 = [\mathbf{h}_{s2}, \mathbf{h}_{t2}, |\mathbf{h}_{s2} - \mathbf{h}_{t2}|, \mathbf{h}_{s2} \odot \mathbf{h}_{t2}] \quad (4)$$

$$y = f_{score}([\mathbf{f}_1, \mathbf{f}_2]) \quad (5)$$

[2] https://wmt-qe-task.github.io/

3.1.2 Task2: Error span detection

Our model was adapted from CometKiwi (Rei et al., 2022). The original binary classification was changed to three-way classification with the following labels: major error, minor error, no error. We disabled the sentence-level prediction head by setting the weight of the original sentence module to 0.

3.2 Corruption-based Data Augmentation

Algorithm 1 Corruption-based data augmentation

Require: source s, translation t, DA $score$
Ensure: $score > 70$

$n \leftarrow min(randint(0, 5), len(t))$
$i \leftarrow 0$
$\hat{t} \leftarrow t$
while $i < n$ **do**
  $\hat{t} \leftarrow corrupt(\hat{t})$
  $i \leftarrow i + 1$
end while
$score_{new} \leftarrow score \times \frac{f_{QE}(s,\hat{t},t)}{f_{QE}(s,t,t)}$
return $s, \hat{t}, score_{new}$

This year's QE shared task primarily focuses on low-resource languages. The scarcity of training data poses a challenge of overfitting. We tried to overcome this problem by augmenting the dataset with various types of noise, including deletion, insertion and substitution errors. Our approach is described in Algorithm 1. We first selected source-translation pairs (s and t) that had a score above 70. We did not use low quality translations for augmentation, as our approach was designed to generate translation with lower scores compared to the original translation. Then, we randomly sampled the number of corruptions and iteratively incorporated these corruptions into the translation, resulting in a new translation ($\hat{t}$). The corruption types are listed as follows:

- **Deletion**: A random word in the translation was deleted.
- **Insertion**: A random word in the translation was selected and inserted in a random position.
- **Substitution**: A random word was replaced with another word in the translation.

To generate a pseudo score for each new translation, we employed a reference-based QE model [3] f_{QE}. The key idea is to use the original translation as the reference and the corrupted translation as the new translation. Since the output of the QE model is in the range between 0 and 1, we can use this value to scale the original score to obtain the pseudo score. However, we observed that even when the reference and translation are the same, the model will not generate a score close to 1, which is inconsistent with the assumption that if there is no corruption, the score should be unchanged. Therefore, we constructed the scaling factor as the ratio between the corrupted translation score and the uncorrupted translation score ($\frac{f_{QE}(s,\hat{t},t)}{f_{QE}(s,t,t)}$). This data augmentation method can be viewed as distilling knowledge from a pre-trained reference-based QE model and it has the potential to increase model generalisability and provide diverse checkpoints for ensemble.

4 Experiments

4.1 Experimental setups

4.1.1 Task1

Our system is built on top of the COMET package [4]. We fine-tuned four pre-trained models, namely XLM-R, InfoXLM, RemBERT and CometKiWi [5], on a single Nvidia Tesla V100 GPU with a batch size of 4, gradient accumulation of 8 and mean square error loss function. We stopped the training when there was no improvement in terms of Spearman correlation on the dev set for five test runs. For each language pair, the augmented dataset, which contains more than ten times data than the original dataset, was pre-generated instead of generated on-the-fly to improve training efficiency. We considered three schedules: training the model with the original dataset; training it with the augmented dataset (only DA training set); and first training it with the augmented dataset and then finetuning it on the original dataset. The training step took around 3 hours and 10 hours with the original and the augmented dataset respectively.

With four base models and three schedules, we obtained twelve checkpoints for each language pair. We ensembled these checkpoints by taking the

[3] https://huggingface.co/Unbabel/wmt22-comet-da
[4] https://github.com/Unbabel/COMET
[5] https://huggingface.co/Unbabel/wmt22-cometkiwi-da

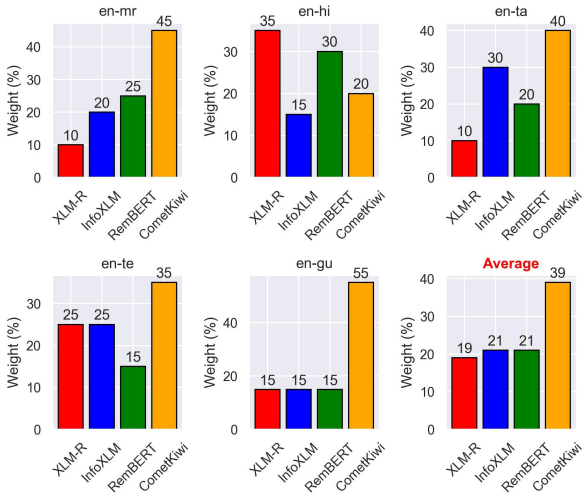


Figure 1: The ensemble weights for each base model.

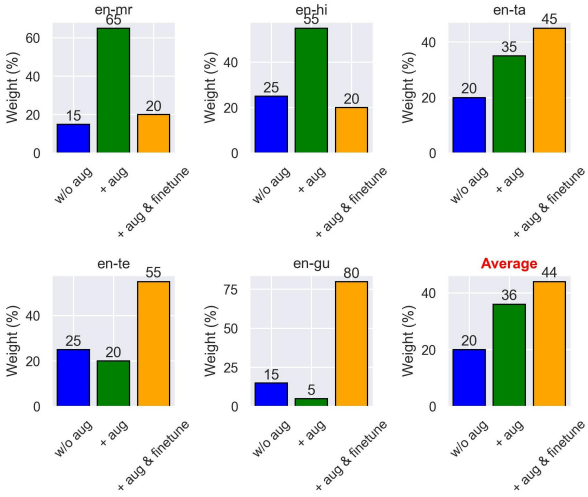


Figure 2: The ensemble weights for different training dataset configurations. 'w/o aug' and '+ aug' mean using the original or augmented dataset respectively. '+ aug & finetune' means training on augmented dataset and then finetuning on the original one.

weighted-average of predicted scores. The weights were optimized using Optuna, an automatic hyperparameter search framework. We used the Spearman correlation as the objective, set the step to 0.05, and ran 1000 trials on the dev set.

4.1.2 Task2

We fine-tuned two pre-trained models, XLM-R and CometKiwi, for 10 epochs with batch size of 32. We created two training subsets using the annotated data from 2020 to 2022. Consequently, four checkpoints were obtained for each language pair. We combined the results of these checkpoints by using the union of the predicted spans, which outperformed token-level majority voting.

4.2 Results

4.2.1 Task1

Results of sentence-level QE in terms of Spearman correlation are shown in Table 1. Without data augmentation, CometKiwi has the best average correlation of 0.597, while XLM-R, InfoXLM and RemBERT are close behind with around 0.585. Figure 1 reveals the importance of each model in the ensemble. CometKiwi has the highest weight for four language pairs, meaning it contributes most to the final prediction. Other base models perform similarly, with XLM-R being most important for en-hi language pair.

The corruption-based data augmentation approach has the most notable benefits for the en-mr language pair. The performance of models based on XLM-R, InfoXLM and CometKiwi are improved significantly. It is worth noting that these models do not need to be fine-tuned on the original training set to achieve comparable or better results than no augmentation, even when more than 90% of the targets are pseudo labels. For other language pairs, data augmentation has limited benefits when used with a single base model. One possible reason is that the reference-based model did not produce high-quality pseudo labels for language pairs with limited resources. However, we did observe that models with data augmentation played important roles in the ensemble. As shown in Figure 2, on average, models without data augmentation were assigned a weight of only 20%, whereas models that were trained purely on augmented data or pretrained on augmented data had a total weight of 80%, indicating that data augmentation can improve the performance of the ensemble and prevent overfitting.

Our final ensemble consists of 12 checkpoints, but some of them have zero weight after optimization. Therefore, the average number of models in the ensemble for each language pair is eight. The ensemble outperforms any single model on the dev set by a noticeable margin. On the test set, the ensemble achieves outstanding results, with Spearman scores higher than 0.69 for three language pairs (en-mr, en-ta, en-gu) and the Spearman of en-ta even reached 0.775. Our submissions are much better than the organizer's baseline. The assessment results of MQM are shown in Table 2. With the model ensemble methods, the assessment results have been significantly improved.

Method	en-mr	en-hi	en-ta	en-te	en-gu	Avg.
XLM-R	0.541	0.614	0.663	0.464	0.644	0.585
+augmentation	0.554	0.613	0.663	0.435	0.608	0.575
+augmentation & finetune	0.554	0.615	0.658	0.442	0.624	0.579
InfoXLM	0.527	0.600	0.663	0.461	0.654	0.581
+ augmentation	0.565	0.607	0.671	0.447	0.635	0.585
+ augmentation & finetune	0.557	0.612	0.669	0.454	0.651	0.589
RemBERT	0.549	0.603	0.663	0.436	0.682	0.587
+ augmentation	0.547	0.587	0.668	0.416	0.622	0.568
+ augmentation & finetune	0.532	0.598	0.659	0.417	0.633	0.568
CometKiwi	0.557	0.598	0.689	0.452	0.689	0.597
+ augmentation	0.580	0.583	0.673	0.458	0.660	0.591
+ augmentation & finetune	0.579	0.588	0.690	0.464	0.677	0.600
Ensemble	**0.592**	**0.636**	**0.707**	**0.481**	**0.699**	**0.623**
baseline **(test set)**	0.392	0.281	0.507	0.193	0.337	0.342
Ensemble **(test set)**	**0.692**	**0.644**	**0.775**	**0.394**	**0.691**	**0.639**

Table 1: Results for sentence-level QE in terms of **Spearman** correlation. Ground-truth annotations were derived from **Direct Assessment**. Except for the last two rows which shows the results on test set, other results were based on the dev set.

Method	en-de	zh-en
XLM-R	0.529	0.293
InfoXLM	0.520	0.213
RemBERT	0.525	0.178
CometKiwi	0.468	0.243
Ensemble	**0.582**	**0.343**
baseline **(test set)**	0.340	0.447
Ensemble **(test set)**	**0.437**	**0.460**

Table 2: Results for sentence-level QE in terms of **Spearman** correlation. Ground-truth annotations were derived from **Multi-dimensional Quality Metrics**.

Method	zh-en	en-de	he-en
XLM-R	0.169	/	/
InfoXLM	0.176	0.143	0.085
+CometKiwi	0.187	0.151	0.095
baseline **(test set)**	0.219	**0.167**	0.227
Ensemble **(test set)**	**0.235**	0.166	**0.266**

Table 3: Results for error span detection in terms of F1 score.

4.2.2 Task2

The results for error span detection are displayed in Table 3. Our system achieved an F1 score of 0.235 on the zh-en language pair, which is significantly higher than the baseline. Moreover, for the language pair without supervised data (he-en), our system achieved a relative improvement of 33% over the baseline.

5 Conclusion

This paper mainly presents HW-TSC's sentence-level QE system called Ensemble-CrossQE. Using our previous year's model CrossQE as the foundation, we carried out comprehensive experiments with different pre-trained models. To further improve the robustness for low-resource language pairs and provide various checkpoints for model ensemble, we introduced a corruption-based data augmentation method. For sentence-level QE task, our system delivers a good performance on all language-pairs with DA annotations. In the future, we will investigate distillation method to transfer the knowledge of the ensemble to a single model to improve efficiency and we plan to leverage external parallel data and translation models for data enhancement. Additionally, in this paper, we only present brief investigations of the error span detection task. Therefore, we plan to further explore word-level QE tasks, which can improve the interpretability of QE.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proc. ACM SIGKDD*, pages 2623–2631.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, He-Yan Huang, and Ming Zhou. 2021. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. In *Proc. NAACL*, pages 3576–3588.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models. *arXiv preprint arXiv:2010.12821*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proc. ACL*, pages 8440–8451.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proc. EMNLP*, pages 2685–2702.

Ricardo Rei, Marcos Treviso, Nuno M Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José GC De Souza, Taisiya Glushkova, Duarte Alves, Luísa Coheur, et al. 2022. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. In *Proc. WMT*, pages 634–645.

Lucia Specia, Carolina Scarton, Gustavo Henrique Paetzold, and Graeme Hirst. 2018. *Quality estimation for machine translation*, volume 11. Springer.

Shimin Tao, Su Chang, Ma Miaomiao, Hao Yang, Xiang Geng, Shujian Huang, Min Zhang, Jiaxin Guo, Minghan Wang, and Yinglu Li. 2022. Crossqe: Hw-tsc 2022 submission for the quality estimation shared task. In *Proc. WMT*, pages 646–652.

Scaling up CometKiwi: Unbabel-IST 2023 Submission for the Quality Estimation Shared Task

Ricardo Rei[*1,2,4], **Nuno M. Guerreiro**[*1,3,4], **José Pombal**[1], **Daan van Stigt**[1], **Marcos Treviso**[3,4], **Luisa Coheur**[2,4], **José G. C. de Souza**[1], **André F. T. Martins**[1,3,4]
[1]Unbabel, Lisbon, Portugal, [2]INESC-ID, Lisbon, Portugal
[3]Instituto de Telecomunicações, Lisbon, Portugal
[4]Instituto Superior Técnico, University of Lisbon, Portugal

Abstract

We present the joint contribution of Unbabel and Instituto Superior Técnico to the WMT 2023 Shared Task on Quality Estimation (QE). Our team participated on all tasks: sentence- and word-level quality prediction (task 1) and fine-grained error span detection (task 2). For all tasks, we build on the CometKiwi-22 model (Rei et al., 2022b). Our multilingual approaches are ranked first for all tasks, reaching state-of-the-art performance for quality estimation at word-, span- and sentence-level granularity. Compared to the previous state-of-the-art, CometKiwi-22, we show large improvements in correlation with human judgements (up to 10 Spearman points). Moreover, we surpass the second-best multilingual submission to the shared-task with up to 3.8 absolute points.

1 Introduction

Quality Estimation (QE) is the task of automatically assigning a quality score to a machine translation output without depending on reference translations (Specia et al., 2018). This paper details the collaborative effort of Unbabel and Instituto Superior Técnico (IST) in the WMT23 Quality Estimation shared task, which encompassed two primary tasks: (i) sentence- and word-level quality prediction and (ii) fine-grained error span detection.

As of last year, some language pairs in the test set were absent from the training data. To address this, following a similar approach to the previous year, our systems were developed to achieve good multilingual generalization and to accommodate previously unseen languages. To achieve this, we start by leveraging the direct assessments (DA) labeled data obtained from the WMT Metrics shared task from 2017 to 2020, the MLQE-PE dataset (Fomicheva et al., 2022), and the training data (DA) specifically annotated for Indian languages in the 2023 shared task edition. In total, these datasets encompass close to 1M annotations covering 38 language pairs. We start by constructing generic models using this corpus. These generic QE models were subsequently fine-tuned for this year's subtasks.

For Task 1 – sentence-level, we fine-tuned our generic models exclusively with this year's DA data. The architecture of these models remains consistent with our submission from the previous year, but we employ XLM-R XL and XXL as pretrained encoders (Conneau et al., 2020). For the word-level quality prediction task, we follow the successful approach of combining the sentence- and word-level signals into one loss during the finetuning step, which has yielded positive results in previous iterations (Rei et al., 2022b). For fine-grained error span detection, we conducted experiments exploring various approaches that build upon our word-level and sentence-level strategies. In terms of contrasting systems, we explored UnbabelQi[1] and GPT-4 (OpenAI, 2023). For GPT-4, we used a prompt designed to predict both the location and severity of errors in each translation, akin to the approach used in AutoMQM (Fernandes et al., 2023).

Overall, our main contributions are: (i) we introduce approaches for multilingual machine translation quality estimation that are consistently first-ranked at word-, span-, and sentence-level granularity; (ii) we explore different approaches to predict the span of problematic translations along with their error severities (OK, Minor, Major); (iii) we publicly release two of our best models for research purposes (CometKiwi -XL[2] and -XXL[3]). To the best of our knowledge, these are the largest open-source QE models publicly released.

Our submitted systems attain the top multilingual results in all tasks: For Task 1 sentence-

[*]Equal contribution. ✉ ricardo.rei@unbabel.com

[1]https://qi.unbabel.com/

[2]https://huggingface.co/Unbabel/wmt23-cometkiwi-da-xl

[3]https://huggingface.co/Unbabel/wmt23-cometkiwi-da-xxl

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 841–848
December 6–7, 2023.

level prediction, our multilingual system achieves 59.4 Spearman correlation points, surpassing the second-best system by nearly 4 absolute points. For word-level, our system achieves a 31.7 MCC score, outperforming the second-best system by almost 2 absolute MCC points. For error span prediction, our multilingual system achieves a 22 F1.0 score, beating the second-best system by more than 5 F_1 points.

2 Overview of the shared-task

QE systems are designed according to the granularity in which predictions are made (e.g., sentence- or word-level QE). In sentence-level QE, the goal is to predict a single quality score $\hat{y} \in \mathbb{R}$ given the whole source and its translation as input. Word-level QE works at a lower granularity level, with the goal of predicting binary quality labels $\hat{y}_i \in \{\text{OK}, \text{BAD}\}$ for all $1 \leq i \leq n$ *machine-translated words*, indicating whether that word is a translation error. In fine-grained error span detection, systems are tasked with flagging which parts of the segment, i.e., sequences of consecutive characters, contain errors. If an error span is found, the system has to point out its severity; in this shared task, an error span's severity can be classified as MINOR or MAJOR. We sometimes refer to the parts of the segment that do not belong to an error span as being labelled as OK. We participated on all tasks of this year's shared-task. We specify the language pairs and the released data below:

Task 1 – Sentence-level quality prediction: Submissions for this task were evaluated based on their correlation with Direct Assessment (DA) annotations for five language pairs: English→Marathi (*en-mr*), English→Hindi (*en-hi*), English→Tamil (*en-ta*), English→Telugu (*en-te*), and English→Gujarati (*en-gu*). Furthermore, they were evaluated using Multidimensional Quality Metrics (MQM) annotations for three language pairs: English→German (*en-de*), Chinese→English (*zh-en*), and English→Hebrew (*he-en*). Training data was made available for all language directions except for *he-en*.

Task 1 – Word-level quality prediction: Submissions for this task underwent evaluation based on tags inferred from post-editions for English→Farsi (*en-fa*) and English→Marathi (*en-mr*). Additionally, they were assessed using MQM annotations for *en-de*, *zh-en*, and *he-en*. No additional training or development data with word-level tags were made available. To the best of our knowledge, no word-level data is available for *en-fa* and *he-en*.

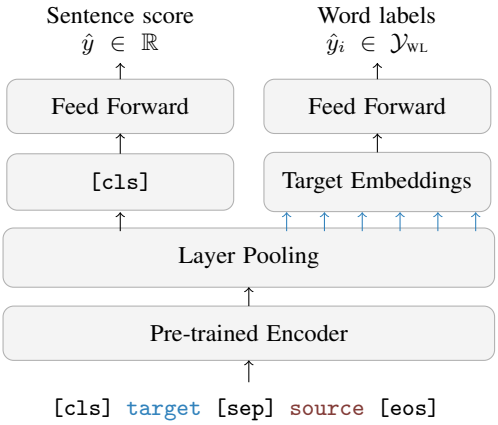


Figure 1: Our model follows COMETKIWI for sentence-level (left part) and word-level QE (right part). We represent the output space of the word-level head by $\mathcal{Y}_{\text{WL}}$.

Task 2 – Fine-grained error span detection: submissions were evaluated on error spans obtained via MQM annotations for 3 language pairs (*en-de*, *zh-en* and *he-en*). No training nor development data is available for *he-en*.

3 Implemented Systems

We largely follow the architecture of COMETKIWI (Rei et al., 2022b) – see Figure 1 for an illustration. We concatenate the machine translated sentence $\boldsymbol{t} = \langle t_1, ..., t_n \rangle$ and its source sentence counterpart $\boldsymbol{s} = \langle s_1, ..., s_m \rangle$ to serve as input to the encoder. This encoder then produces hidden state matrices $\boldsymbol{H}_0, ..., \boldsymbol{H}_L$ for each layer $0 \leq \ell \leq L$, where $\boldsymbol{H}_\ell \in \mathbb{R}^{(n+m)\times d}$, where $\ell = 0$ corresponds to the embedding layer and d is the hidden size. Following this, all hidden states are fed to a scalar mix module (Peters et al., 2018) that learns a weighted sum of the hidden states of each layer of the encoder, producing a sequence of aggregated hidden states $\boldsymbol{H}_{\text{mix}}$ as follows:

$$\boldsymbol{H}_{\text{mix}} = \lambda \sum_{\ell=0}^{L} \beta_\ell \boldsymbol{H}_\ell. \quad (1)$$

Here λ is a scalar trainable parameter, $\boldsymbol{\beta} \in \triangle^L$ is given by $\boldsymbol{\beta} = \mathsf{sparsemax}(\boldsymbol{\phi})$ using a sparse transformation (Martins and Astudillo, 2016), with $\boldsymbol{\phi} \in \mathbb{R}^L$ as learnable parameters, and where we

denote by $\triangle^L := \{\boldsymbol{\beta} \in \mathbb{R}^L : \mathbf{1}^\top \boldsymbol{\beta} = 1, \boldsymbol{\beta} \geq 0\}$ the probability simplex.[4]

For sentence-level models, we use the hidden state of the `<cls>` token as the sentence representation, which, in turn, is passed to a 2-layered feed-forward module in order to get a sentence score prediction $\hat{y} \in \mathbb{R}$. For word-level and error span detection models, we first retrieve the hidden state vectors associated with each each token in $\boldsymbol{t}$, and then pass them to a linear projection to get word-level predictions $\hat{y}_i \in \mathcal{Y}_{\text{WL}}$, $\forall_{1 \leq i \leq n}$. The output space of the word-level predictions is different depending on whether the models are constructed for word-level quality prediction ($\mathcal{Y}_{\text{WL}} = \{\text{OK}, \text{BAD}\}$), or error span detection ($\mathcal{Y}_{\text{WL}} = \{\text{OK}, \text{MINOR}, \text{MAJOR}\}$).

Pretrained multilingual encoders. Similarly to (Rei et al., 2022b), we employ InfoXLM L (Chi et al., 2021).[5] Additionally, we experiment with scaled-up multilingual encoders, including XLM-R XL,[6] and XLM-R XXL.[7] InfoXLM L comprises 24 encoder blocks with 16 attention heads each, totaling 550M parameters. XLM-R XL and XLM-R XXL have 32 attention heads for each encoder block, 36 and 48 encoder blocks and a total of 3.5B and 10.7B parameters, respectively.

Generic models for all tasks. We create, for each model size, a generic model that will then be further adapted to each separate task. To train these models, we use the collective corpora from 2017 to 2019 DA annotations of the WMT Translation shared task, and the MLQE-PE corpus (Fomicheva et al., 2022). We include the human annotations respective to the language pairs of this year's shared task for 7 different language pairs: DA annotations for *en-mr*, *en-hi*, *en-ta*, *en-te*, *en-gu*, and MQM annotations for *en-de* and *zh-en*. Overall, the generic models are trained on sentence-level quality prediction with over 940k samples with source, translation and quality score on 38 different language pairs. When presented with multiple DA scores for the same sentence pair, we used the z-score of the DAs for training but we first normalize the DAs between 0 and 1, where 1 represents a perfect translation and 0 a random one.

Task adaptation. After having obtained the generic models, we will train models for each separate stream of the shared-task, i.e., sentence-level, word-level or error span prediction. To do so, we consider the multi-task optimization from Rei et al. (2022b) wherein sentence scores can be used alongside supervision from word-level tags. Formally,

$$\mathcal{L}_{\text{SL}}(\theta) = \frac{1}{2}(y - \hat{y}(\theta))^2 \tag{2}$$

$$\mathcal{L}_{\text{WL}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n} w_{y_i} \log p_\theta(y_i) \tag{3}$$

$$\mathcal{L}(\theta) = \lambda_{\text{SL}}\mathcal{L}_{\text{SL}}(\theta) + \lambda_{\text{WL}}\mathcal{L}_{\text{WL}}(\theta), \tag{4}$$

where $w \in \mathbb{R}^{|\mathcal{Y}_{\text{WL}}|}$ represents the class weights given for the word-level tags,[8] and $\lambda_{\text{SL}}, \lambda_{\text{WL}} \in \mathbb{R}_+$ are used to weigh the sentence and word-level losses, respectively. Note that $\lambda_{\text{SL}} = 1$ and $\lambda_{\text{WL}} = 0$ yields a fully sentence-level model, whereas $\lambda_{\text{SL}} = 0$ and $\lambda_{\text{WL}} = 1$ yields a word-level model.

Using unconstrained models. For error span detection, we evaluate UnbabelQi, an Unbabel demo QE system, alongside GPT4 (OpenAI, 2023). We prompt GPT4 to produce an MQM annotation for each source-target pair, based on five-shot examples which vary across language pairs but are consistent within segments of the same language pair. We also apply this system in Task 1, deriving a sentence-level score from error spans, in alignment with the MQM framework. This approach bears similarity to AutoMQM (Fernandes et al., 2023).

3.1 Task 1: Quality prediction

After the pretraining phase, we *further* **separately** adapt the generic models to the released DA and MQM data for this year's shared task.

3.1.1 Sentence-level quality prediction

Adaptation for Sentence-level. In order to tailor our models to the language pairs featured in this year's shared task, we conducted full fine-tuning until convergence on the released validation set. This fine-tuning exclusively leveraged the recently released Direct Assessment (DA) annotations for this year's task. This approach yields additional improvements for those languages. In the case of the MQM language pairs, our preliminary experiments revealed that attempting significant performance

[4] As it has been shown in (Rei et al., 2022a) not all layers are relevant and thus, using `sparsemax` we learn to ignore layers that do not help in the task at hands.

[5] https://huggingface.co/microsoft/infoxlm-large

[6] https://huggingface.co/facebook/xlm-roberta-xl

[7] https://huggingface.co/facebook/xlm-roberta-xxl

[8] These parameters help control how much we penalize the different granularities of word-level errors.

improvements on the MQM data led to noteworthy drops in correlations for the other language pairs using DAs. Consequently, for the MQM language pairs, we opted to employ the generic models as they are.

Ensembling models. Similarly to Rei et al. (2022b), we use Optuna (Akiba et al., 2019) to assemble four models – two XL and two XXL – into a single system. We do so by finding the optimal weights for each language pair among these four multilingual models, and combining their predictions according to those weights. Notably, the XXL models are generic models, whereas the two XL checkpoints were further optimized with this year's shared task data. As expected, the XL models carry more weight for Indian languages, while the XXL generic models were deemed more crucial for MQM languages.

3.1.2 Word-level quality prediction

For the word-level QE tasks, we experimented with both the multi-task setting and word-labels only.

Training word-level models. This year, no training or development data with word-level tags were made available. As such, the training data for our models consists of the training data used in Rei et al. (2022b), combined with the development sets from the 2022 WMT Shared Task. As the word-level task was going to be tested in a zero-shot scenario for two out of five language pairs (*en-fa*, *he-en*), contrary to Rei et al. (2022b), we do not prepend a language prefix to the beginning of the source and target segments during training. Moreover, for the post-edit (PE) models, we removed samples from two language pairs (*ps-en* and *en-cs*) from the training data. We did so to assess, during validation, the models' capability to generalise in a zero-shot scenario. For the MQM models, we used all available annotations, including those in *en-ru*.

Ensembling models. For word-level we followed a similar ensembling technique used for sentence-level. Specifically, we combined multiple systems trained with different hyperparameters, encoder size and pre-training setups. In the case of word-level predictions, we aggregate multiple predictions into OK/BAD tags by following the *ensemble-tags* procedure from Rei et al. (2022b). In this approach, we combine the predicted tags of each model: for every input segment, we get a combined tag, $\alpha \sum_{i \in \mathcal{M}} w_i c_i$, where c_i is the tag predicted by the model and α is the weight for the BAD tag. We use Optuna to determine the optimal weights w_i for each model and the optimal BAD weight α for each LP. In the final submission, we combine six models (five PE models and one MQM model). Five of these models use InfoXLM as the encoder model, and one PE model uses XLM-R XL.[9] Refer to Table 2 for the test set results.

3.2 Task 2: Fine-grained error span detection

In this task, we investigated three distinct approaches. The first approach extends word-level models by modifying their output predictions. More precisely, it involves transforming consecutively predicted BAD tags into character-level error spans, rather than categorizing individual words based on the first subword. To determine the error severities of these spans, we considered two options: labeling all the subwords within the span as either MINOR or MAJOR. Our best results were achieved with the latter approach.

The second approach leverages XCOMET (Guerreiro et al., 2023) in conjunction with a pseudo-reference obtained from DeepL or Google Translate.[10] Similar to our models from Task 1 word-level, XCOMET is trained with a multitask objective. Additionally, XCOMET is simultaneously optimized for both reference-free and reference-based evaluation, following UNITE (Wan et al., 2022). During inference, XCOMET can leverage a reference translation to enhance error identification. Since we employ a pseudo-reference that may contain translation errors, we initially assess the quality of the pseudo-reference using a generic QE system from Task 1 (reference_score). For all pseudo-references with a score below 0.5, we run XCOMET with QE-only input. For pseudo-references scoring above 0.5, the input weights for XCOMET are determined as follows:

$$\begin{aligned} \text{diff} &= 1 - \text{reference\_score} \\ \text{src\_weight} &= 2 \cdot \text{diff} \\ \text{ref\_weight} &= (1 - \text{src\_weight}) \cdot 0.4 \\ \text{uni\_weight} &= (1 - \text{src\_weight}) \cdot 0.6 \end{aligned}$$

Here, `src_weight` represents the weight assigned to the source-only input, `ref_weight` de-

[9] We found it hard to obtain performance boosts by scaling up to XLM-R XL on the word-level task. As such, we did not experiment with XLM-R XXL.

[10] We choose the best translation using the generic XXL model from task 1.

	DA					MQM			
Encoder	**en-mr**	**en-hi**	**en-ta**	**en-te**	**en-gu**	**en-de**	**zh-en**	**he-en**†	**avg.**
2nd place (Yan et al., 2023)									0.556
CometKiwi-22 (Rei et al., 2022b)									
InfoXLM L	0.625	0.394	0.549	0.229	0.577	0.413	0.476	0.619	0.485
Generic models									
InfoXLM L	0.661	0.505	0.641	0.282	0.661	0.422	0.448	0.610	0.529
XLM-R XL	0.664	0.536	0.607	0.335	0.637	0.422	0.469	0.624	0.537
XLM-R XXL	0.685	0.520	0.670	0.326	0.655	0.443	0.476	0.662	0.555
Further adapted models for sentence-level									
XLM-R XL	0.684	0.583	0.682	0.386	0.683	0.434	0.441	**0.696**	0.574
XLM-R XXL	0.693	0.555	0.738	0.359	0.701	0.434	0.457	0.661	0.575
Final Ensemble									
Ensemble 4x	**0.702**	**0.598**	**0.739**	**0.389**	**0.714**	**0.448**	**0.493**	0.668	**0.594**
GPT4-based model									
GPT4-QE	0.379	0.212	0.146	0.174	0.297	0.442	0.412	0.488	0.319

Table 1: Results for sentence-level QE in terms of Spearman correlation. We represent zero-shot LPs with †.

notes the typical metric input (reference-only input), and `uni_weight` represents a unified input where the model receives all three sentences (translation, source, and reference). Notably, for pseudo-references with a QE score of 1, we rely solely on a reference-only input and the unified input. We refer to this approach as XCOMET-PS-REF.

We also contrast the aforementioned approaches with two unconstrained QE systems: UnbabelQi and GPT-4, as mentioned in Section 3. We refer to these approaches as UNBABELQI and GPT4-QE, respectively.

4 Experimental Results

We present the results on the official test set for each of the tasks for multiple model/data configurations. Sentence-level submissions were evaluated using the Spearman rank correlation. Pearson and Kendall correlation were also used as secondary metrics, but here we report only Spearman since it was the primary metric used to rank systems. word-level submission were evaluated using MCC, F_1-OK, and F_1-BAD, but we report only MCC as it was considered the main metric. Error span detection was evaluated using F_1 score in which the positive labels are all the characters belonging to erroneous spans. Furthermore, each true positive is downweighted to half if the system failed to classify the error span's severity (e.g., MINOR instead of MAJOR). The submitted systems were independently evaluated on in-domain and zero-shot LPs for direct assessments and MQM.

4.1 Quality Estimation

Sentence-level. Results for sentence-level are presented in Table 1. Results indicate that retraining the system from the previous year, specifically COMETKIWI with InfoXLM, using data that encompasses this year's DA, leads to significant improvements. Remarkably, this improvement in correlations is achieved while maintaining the same level of correlations for *en-de* (a high-resource language pair for which both models share the same data) and *he-en*, a language pair that both models had not seen during training. Surprisingly, there was a drop in correlations for *zh-en* even though both models saw the same *zh-en* data. Nevertheless, the overall performance of the newly retrained version improved by 4.4 Spearman points.

As anticipated, among the three backbone transformers, the XXL model is the top performer, with significant improvements across all language pairs when compared to InfoXLM. Moreover, additional finetuning on this year's training data results in further improvements for the Indian languages. Notably, concerning the MQM data, this supplementary finetuning step not only preserves performance but sometimes even increases it. Similar to last year, the ensemble of high-performing models once again makes up our best submission.

Finally, despite performing well in Task 2, GPT4-QE shows poor correlations at sentence-level prediction with the exception of the *en-de* for which GPT4-QE, although lagging behind the ensemble approach, surpasses our individual models.

Method	Post-edit		MQM			
	en-fa†	en-mr	en-de	zh-en	he-en†	avg.
Baseline (Rei et al., 2022b)	0.293	0.287	0.179	0.225	0.275	0.226
2nd place (Yan et al., 2023)						0.298
Adapted models for word-level						
PE model (InfoXLM L)	0.343	0.343	0.227	0.253	0.382	0.310
PE model (XLM-R XL)	0.325	0.344	**0.255**	0.197	0.306	0.285
MQM model (InfoXLM L)	0.296	0.252	0.215	0.269	0.334	0.273
Final Ensemble						
Ensemble PE + MQM	**0.345**	**0.347**	0.246	**0.302**	**0.402**	**0.317**

Table 2: Results for word-level QE in terms of MCC for the post-edit and MQM LPs. The ensemble is composed by multiple post-edit and MQM models. We represent zero-shot LPs with †.

Word-level. We report the best individual systems Table 2. Our best individual systems were trained on top of the InfoXLM L generic model. For PE models, we used multi-task objective in Eq. 4, as we found that combining the sentence-level and word-level loss was beneficial. However, for MQM models, we trained word-level only models, by setting $\lambda_{\text{SL}} = 0.0$ and $\lambda_{\text{WL}} = 1.0$.

Interestingly, we found that PE models are very competitive on MQM language pairs. For example, the best overall performance for *he-en* was actually obtained with a PE word-level model. This is also reflected on the Optuna weights obtained for our final ensemble, wherein the weights of the PE models are significantly higher than those of the MQM models for all language pairs but *en-de*. In fact, our final ensemble for *en-zh* and *en-he* consists solely of PE models trained with different learning rates, λ_{SL}, λ_{WL} and w. Further investigation on two different vectors may lead to improved word-level models: (i) balancing DA and MQM word-level annotations, and (ii) appropriately leveraging the larger capacity of scaled up encoder models.

Fine-grained error span detection. Results for fine-grained error span detection are shown in Table 3. Using a word-level model to obtain error span predictions leads to reasonable performance, comparable to our unconstrained submission, UNBABELQI, a model directly tasked with error span detection. That said, XCOMET-PS-REF, an error span detection model, surpassed both of the previous approaches. We attribute the improved performance to this system being an ensemble of two significantly larger models, and to the usage of a pseudo-reference. We found the latter to be particularly beneficial on *he-en*, a language pair for which we had no training data.

Method	en-de	zh-en	he-en†	avg.
2nd place (Li et al., 2023)				0.165
Baseline	0.167	0.219	0.083	0.156
WORD-LEVEL	0.235	**0.272**	0.105	0.204
XCOMET-PS-REF	0.259	0.270	**0.125**	0.218
UNBABELQI	0.249	0.227	0.111	0.196
GPT4-QE	**0.273**	0.265	0.121	**0.220**

Table 3: Results for fine-grained error span detection (Task 2). Evaluation metric is F_1 score. We represent zero-shot LPs with †. The first two systems are constrained while the other two are unconstrained submissions.

The best approach in terms of average F_1 was GPT4-QE, mostly due to the improved performance on *en-de*. While this is a promising finding for LLM-based quality estimation systems, there are limitations. First, obtaining a sentence-level score from the error spans (as per the MQM framework) leads to poor correlations with human judgements derived from DA (see Table 1) and with low-resource language-pairs like *he-en*. Second, despite being useful in practice and leading to gains in F_1, it is hard to control GPT's precision and recall. We found that the number of examples included in the prompt, their ordering, and the number of errors within each example led to noticeable changes in the system's propensity to flag errors. Thirdly, running QE with a system such as GPT-4 is expensive and slow even for a shared task exercise.

5 Final Remarks

We describe Unbabel and IST joint submission to WMT23 QE shared task. Our approaches correlate well with human judgements for all the three granularities of translation quality prediction, ranking first in all multilingual tasks and surpassing the pre-

vious state-of-the-art model, COMETKIWI-22, by up to 10 Spearman correlation points. Overall, our models follow the same architecture of last year's participation, COMETKIWI. However, this year we leverage more data and larger encoder models. Our best final systems are ensembles of different models trained on DA, post-edits or MQM scores that complement each other. Interestingly, our best systems surpass GPT-4 by a large margin for sentence-level translation quality prediction, and they are comparable to GPT-4 at error span detection.

Acknowledgements

This work was supported by the Portuguese Recovery and Resilience Plan (PRR) through project C645008882-00000055, Center for Responsible AI, by the European Research Council (DECOLLAGE, ERC-2022-CoG 101088763), by EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), and by the Fundação para a Ciência e Tecnologia (contracts UIDB/50021/2020 and UIDB/50008/2020).

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André F. T. Martins, Graham Neubig, Ankush Garg, J. Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. *ArXiv*, abs/2308.07286.

Marina Fomicheva, Shuo Sun, Erick Fonseca, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2022. MLQE-PE: A Multilingual Quality Estimation and Post-Editing Dataset. In *Proceedings of the Language Resources and Evaluation Conference*, pages 4963–4974, Marseille, France. European Language Resources Association.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023. XCOMET: Transparent machine translation evaluation through fine-grained error detection.

Yuang Li, Chang Su, Ming Zhu, Xinglin Piao, Mengyao Lyu, Min Zhang, and Hao Yang. 2023. HW-TSC 2023 Submission for the Quality Estimation Shared Task. In *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Andre Martins and Ramon Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International Conference on Machine Learning*, pages 1614–1623.

OpenAI. 2023. GPT-4 technical report.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, José G.C. de Souza, Pedro G. Ramos, André F.T. Martins, Luisa Coheur, and Alon Lavie. 2022a. Searching for COMETINHO: The little metric that could. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 61–70, Ghent, Belgium. European Association for Machine Translation.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. 2018. Quality estimation for machine translation. *Synthesis Lectures on Human Language Technologies*, 11(1):1–162.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 2022.

UniTE: Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127, Dublin, Ireland. Association for Computational Linguistics.

Zeyu Yan, Wenbo Zhang, Qiaobo Deng, Hongbao Mao, Jie Cai, and Zhengyu He. 2023. IOL Research's Submission for WMT 2023 Quality Estimation Shared Task. In *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

SurreyAI 2023 Submission for the Quality Estimation Shared Task

Archchana Sindhujan[*1], Diptesh Kanojia[1], Constantin Orasan[2] and Tharindu Ranasinghe[3]
[1] Institute for People-Centred AI, University Of Surrey , UK
[2]Centre for Translation Studies, University of Surrey, UK
[3] School of Computer Science and Digital Technologies, Aston University, UK
{a.sindhujan,d.kanojia,c.orasan}@surrey.ac.uk,
t.ranasinghe@aston.ac.uk

Abstract

Quality Estimation (QE) systems are important in situations where it is necessary to assess the quality of translations, but there is no reference available. This paper describes the approach adopted by the SurreyAI team for addressing the Sentence-Level Direct Assessment shared task in WMT23. The proposed approach builds upon the TransQuest framework, exploring various autoencoder pre-trained language models within the MonoTransQuest architecture using single and ensemble settings. The autoencoder pre-trained language models employed in the proposed systems are XLMV, InfoXLM-large, and XLMR-large. The evaluation utilizes Spearman and Pearson correlation coefficients, assessing the relationship between machine-predicted quality scores and human judgments for 5 language pairs (English-Gujarati, English-Hindi, English-Marathi, English-Tamil and English-Telugu). The MonoTQ-InfoXLM-large approach emerges as a robust strategy, surpassing all other individual models proposed in this study by significantly improving over the baseline for the majority of the language pairs.

1 Introduction

The primary objective of quality estimation (QE) systems is to assess the quality of a translation without relying on a reference translation. This make QE valuable within translation processes, as it enables the determination of whether an automatically generated translation is sufficiently accurate for a specific purpose. This aids in deciding whether the translation can be used as is, requires human intervention for full translation, or necessitates post-editing by a human translator (Kepler et al., 2019b). Quality estimation can be conducted across various levels: word/phrase level, sentence level and document level. This paper considers only the sentence-level QE and presents our participation in the WMT23 Sentence-level direct assessment (DA) shared task. In the context of this task, participating systems are required to predict the DA score for a given (source, target) pair. This score serves as a measure of the translation quality.

Building upon the ideas presented in TransQuest by Ranasinghe et al. (2020b), our investigation explores the use of various pre-trained models within the MonoTransQuest architecture for the sentence-level quality estimation shared task. The architecture employs autoencoder pre-trained language models to fine-tune the QE data to predict a score which indicates the quality of translation. Using the MonoTransQuest architecture as the base we employ the pre-trained transformers separately to implement the systems MonoTQ-XLMV, MonoTQ-InfoXLM-large and MonoTQ-XLMR-large. In addition, we propose ensembleTQ which combines the output of MonoTransQuest when using different pre-trained models. All the proposed systems achieve a significantly higher Spearman correlation score compared to the baseline.

The paper is structured as follows. Section 2 briefly presents related work on quality estimation. Section 3 provides a concise overview of the dataset used in the sentence-level QE shared task. Moving on to Section 4, we introduce the autoencoder pre-trained language models and proposed systems and detail the training methodology. Section 5 is dedicated to the evaluation and Section 6 comprises the result and discussion. The paper concludes by summarizing findings, highlighting conclusions, and suggesting potential avenues for future research in the final section.

2 Related work

Quality estimation in machine translation has evolved significantly throughout the years. Initially, it relied on feature engineering and conventional machine learning techniques like SVM and basic

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 849–855

neural networks (Specia et al., 2015; Scarton and Specia, 2014). However, Neural Networks has since become central to quality estimation, where there is no more need of feature engineering, and the models can be trained directly on the data (Kepler et al., 2019b,a; Specia et al., 2018). Recently, Transformer-based architectures have arisen as robust solutions for machine translation and quality estimation. Notably, there are two widely recognized frameworks that leverage this transformative approach for QE tasks: TransQuest (Ranasinghe et al., 2020a) and CometKIWI (Rei et al., 2022).

Ensemble methods have also been explored extensively in Quality Estimation tasks (Bao et al., 2022; Geng et al., 2022; Kepler et al., 2019b; Ranasinghe et al., 2020b; Rei et al., 2022). The ensemble approach from Lim and Park (2022) using K-folds consistently outperformed the standard method, underscoring the prevalent belief that ensemble strategies enhance performance outcomes. However, some of the research studies (Ranasinghe et al., 2020b; Bao et al., 2022; Rei et al., 2022) show that combining multi-lingual models through ensembling yields better results than the traditional k-fold ensemble technique. Geng et al. (2022) suggest an alternative ensemble method that merges the results from models trained using various sentence-level metrics.

Our study delves into the performance of cutting-edge pre-trained transformer-based approaches when applied to sentence-level Quality Estimation tasks.

3 Dataset

We focus on Sentence-Level Direct Assessment tasks which comprise datasets for 5 language pairs which has English on the source side and Indian languages on the target side: English-Gujarati (En-Gu), English-Hindi (En-Hi), English-Marathi (En-Mr), English-Tamil (En-Ta) and English-Telugu (En-Te). Among these language pairs, En-Hi language pair is considered mid-resourced and all the other language pairs are low-resourced. Each language pair includes around 7,000 sentence pairs in the training set, as well as around 1,000 sentence pairs in both the development and testing sets. Each translation was evaluated by three professional translators who assigned a score between 0 and 100. These Direct Assessment (DA) scores were normalized using the z-score. The final score for the sentence-level task requires predicting the mean DA z-scores for the test sentence pairs. More details on this can be found in Zerva et al. (2022).

4 Methodology

This section outlines the approach taken to formulate our quality estimation techniques. We begin by detailing the autoencoder pre-trained language models employed in our architecture. Then we explain the architecture and the strategy employed to train these network architectures in detail.

4.1 Pre-trained models for fine-tuning

1. XLMR-large

   XLM-Roberta (Conneau et al., 2020) is a pre-trained transformer-based language model which is a part of the Cross-lingual Language Model (XLM). This model employs large-scale cross-lingual pre-training to capture contextual information and representations across 100 languages. The model is trained on 2.5TB of filtered CommonCrawl data from multiple languages, allowing it to effectively learn cross-lingual and language-specific patterns. The XLM-R architecture takes sequences as input, with a maximum token limit of 512, and generates contextualized embeddings for each token, enabling it to perform well on various natural language processing tasks across different languages (Ranasinghe et al., 2020b, 2021).

2. XLMV

   XLMV is a multilingual language model with a one million token vocabulary trained on 2.5TB of data from Common Crawl (same as XLM-R) (Liang et al., 2023). In the context of large multilingual language models, a common practice involves employing a single vocabulary shared across a diverse set of languages. Even with the expansion in model complexity, including parameter count and depth, the vocabulary size has remained relatively static. This constraint in vocabulary hampers the potential of multilingual models such as XLM-R to capture nuanced representations effectively (Wang et al., 2019).

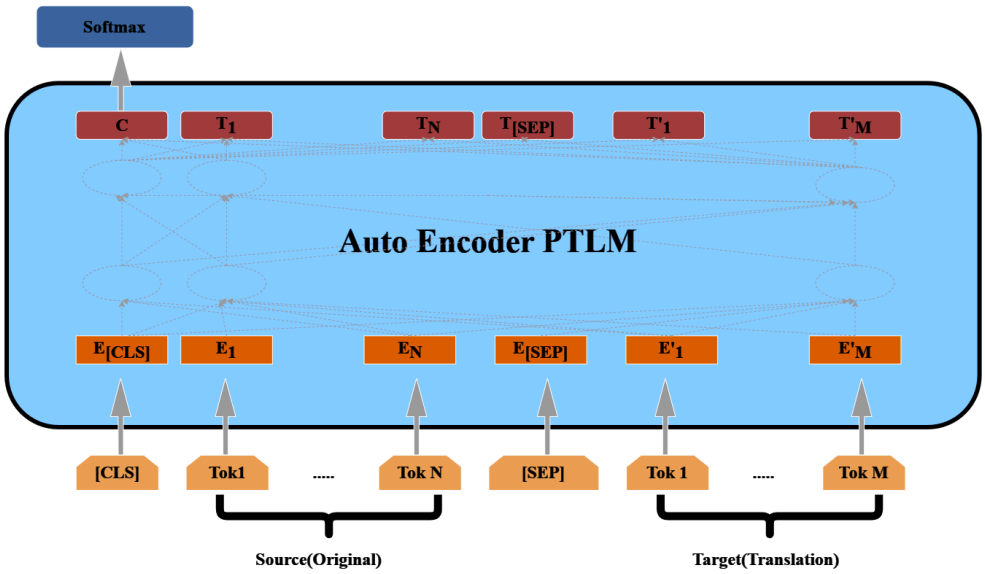


Figure 1: Architecture diagram of the proposed approaches

XLMV introduced an innovative strategy addressing this issue by achieving scalability to extensive multilingual vocabularies. XLMV involves prioritizing vocabulary allocation based on language-specific lexical overlap, ensuring sufficient coverage for each language. The outcome is tokenizations that hold enhanced semantic significance and are generally more concise compared to those generated by XLM-R.

3. InfoXLM-large

   InfoXLM-large (Chi et al., 2021), is an information-theoretic framework for cross-lingual language model pre-training. It extends the XLM-R architecture by formulating cross-lingual pre-training to maximize mutual information between multilingual texts at different granularities. This approach enhances the model's capability to learn effective cross-lingual representations by capturing shared information across languages. InfoXLM-large introduces a novel pre-training task based on contrastive learning, treating bilingual sentence pairs as views of the same meaning. By jointly training on monolingual and parallel corpora, the model improves the transferability of its representations for various downstream cross-lingual tasks (Rei et al., 2022; Bao et al., 2022).

4.2 Architecture

The proposed architecture of MonoTransQuest employs a pre-trained language model as shown in Figure 1. The MonoTransQuest architecture in TransQuest (Ranasinghe et al., 2020b) considers only the XLMR transformer model. In our proposed system, we train multiple multilingual QE models by fine-tuning autoencoder pre-trained language models (PTLMs) and report mean z-scores. The PTLMs are namely XLMV, InfoXLM-large, and XLMR-large which we have explained in section 4.1. The model's input consists of the original sentence (source) and its translation (target) concatenated, with a [SEP] token. This token marks the separation of the original sentence and the translated sentence. The pre-trained auto-encoder accepts input sequences with a token limit of 512 and produces a sequence representation as output. The initial token of the sequence is [CLS] token, encompassing a distinctive embedding to signify the entire sequence. Subsequently, embeddings are assigned to each word in the sequence. Ranasinghe et al. (2020b) highlights the superiority of the CLS-strategy over the MEAN-strategy (calculating the mean of all output vectors corresponding to the input words) and MAX-strategy (determining the maximum value across the output vectors of input words) for pooling within the MonoTransQuest framework. We have used the CLS-strategy (using the output of the [CLS] token) to extract the output from the transformer model. Consequently, we employed

	Method	En-Gu		En-Hi		En-Mr		En-Ta		En-Te	
		ρ	r	ρ	r	ρ	r	ρ	r	ρ	r
I	Baseline	0.337	0.307	0.281	0.245	0.392	0.427	0.507	0.402	0.193	0.153
II	MonoTQ-XLMV	0.673	0.536	0.572	0.687	**0.642**	0.425	0.670	0.559	**0.464**	0.642
III	MonoTQ-InfoXLM-large	**0.713**	0.656	**0.624**	0.726	0.470	0.030	**0.726**	0.662	0.462	0.719
IV	MonoTQ-XLMR-large	0.438	0.299	0.440	0.430	0.395	-0.117	0.482	0.454	0.345	0.211
V	ensembleTQ	0.649	0.700	0.551	0.668	0.596	0.668	0.674	0.710	0.349	0.376

Table 1: Spearman (ρ) and Pearson (r) correlation between the proposed approach predictions and human DA judgments. The best Spearman score obtained for each language pair (any method) is marked in bold. Rows II, III, and IV indicate the single-configuration settings of MonoTransQuest architecture with different pre-trained transformer models as explained in Section 5.1, and ensembleTQ in row V is explained in Section 5.2. The baseline results are in Row I.

the [CLS] token's embedding as input for a softmax layer. The softmax layer predicts the translation's quality score. The mean-squared-error loss function was used as the objective function for training.

4.3 Training and Implementation Details

We started the training with MonoTQ-XLMV which incorporates the XLMV-base model with MonoTransQuest for all 5 language pairs. We had the batch size as 8. We have used Adam Optimizer (Kingma and Ba, 2014) with a learning rate of 2e-5. The model is trained using 3 epochs. The training process exclusively utilized the training data. Early stopping was enforced if the evaluation loss failed to show improvements over ten consecutive evaluation rounds. We continued the training with the same set of configurations for MonoTQ-InfoXLM-large and MonoTQ-XLMR-large separately. MonoTQ-XLMV and MonoTQ-InfoXLM-large required twice the training time compared to MonoTQ-XLMR-large which required approximately 40 minutes of training on a GPU with 48GB of memory.

The proposed systems are built upon the most up-to-date version of `TransQuest`[1] framework and executed using Python 3.9 and PyTorch 2.0.1. The integration of pre-trained encoders (`XLMV`[2], `XLMR-large`[3] and `InfoXLM-large`[4]) into the MonoTransQuest architecture was facilitated through the application of HuggingFace's Transformers library.

[1]https://github. com/tharindudr/transQuest
[2]https://huggingface.co/facebook/xlm-v-base
[3]https://huggingface.co/xlm-roberta-large
[4]https://huggingface.co/microsoft/infoxlm-large

5 Evaluation

In this section, we outline the evaluation outcomes of our models. We assess the performance of the proposed models under two circumstances: single model configuration and ensembleTQ.

The primary evaluation criterion employed was Spearman's rank correlation coefficient (Sedgwick, 2014), which is a statistical measure used to evaluate the strength and direction of association between two variables. Also, we have calculated the Pearson correlation coefficient (Cohen et al., 2009) as a secondary metric for the evaluation. In the context of Quality Estimation (QE) for machine translation, it is used to evaluate the correlation between the machine-predicted quality scores and the gold standard labels provided by human annotators in the test dataset. Spearman's rank correlation coefficient assesses the monotonic relationship between the two variables, unlike the Pearson correlation (Cohen et al., 2009), which measures the linear relationship between two variables. It is calculated by first ranking the values of both variables in ascending or descending order and then computing the Pearson correlation coefficient between the two sets of ranks. Spearman's rank correlation coefficient is often preferred because it is less sensitive to outliers and non-linear relationships between the predicted scores and human scores.

		En-Gu		En-Hi		En-Mr		En-Ta		En-Te	
	Team	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r
1	Unbabel-IST	**0.714**	**0.745**	0.598	0.667	**0.704**	**0.735**	0.739	0.733	0.388	0.362
2	IOL Research	0.695	0.742	0.6	0.667	0.505	0.372	0.74	0.742	0.376	0.344
3	HW-TSC	0.691	0.714	**0.644**	**0.72**	0.692	0.718	**0.775**	**0.778**	**0.394**	0.35
4	MMT	0.54	0.581	0.494	0.57	0.65	0.663	0.547	0.531	0.337	0.281
5	Baseline	0.337	0.307	0.281	0.245	0.392	0.427	0.507	0.402	0.193	0.153
6	SurreyAI-ensembleTQ	0.649	0.700	0.551	0.668	0.596	0.668	0.674	0.710	0.349	**0.376**

Table 2: Spearman (ρ) and Pearson (r) correlation between the predictions from the participated systems in WMT23 sentence-level QE shared task and human DA judgments. The best Spearman and Pearson score obtained for each language pair is marked in bold. Even though we have experimented with the single model configurations, we only submitted our ensembled approach (SurreyAI-ensembleTQ) for the shared task competition.

5.1 Single model configurations

Initially, our evaluation focused on the single model configurations of the proposed framework. This involved training a quality estimation model using a single autoencoder pre-trained language model on the training data for each language pair separately. Subsequently, we assessed each model's performance (MonoTQ-XLMV, MonoTQ-InfoXLM-large, MonoTQ-XLMR-large) on the corresponding test set for each language pair. The outcomes of this evaluation for the single model configuration are presented in Table 1.

5.2 EnsembleTQ

Recently, ensemble techniques have demonstrated their efficacy in enhancing transformer-based models' performance (Xu et al., 2020). Following this approach, we employed an ensemble strategy to experiment further to see whether it enhance the performance. For every input within the test set, we aggregate the output scores from various distinct pre-trained models integrated into the MonoTransQuest architecture. Subsequently, we calculate the average of the cumulative score, divided by the number of pre-trained models, resulting in the ensembleTQ score. Finally, we compute the Spearman and Pearson correlation scores for the ensembleTQ score, providing a comprehensive evaluation of our ensemble approach.

6 Result and Discussion

The research is divided into two distinct settings, as outlined in Sections 5.1 and 5.2. The primary evaluation metric employed in this study is the Spearman correlation coefficient.

As shown in Table 1, is notable that the baseline model does not surpass our proposed approaches in terms of Spearman correlation scores in most cases. This outcome underscores the specific strengths and limitations associated with different model architectures. Complementing the Spearman correlation analysis, the examination of Pearson correlation scores further enriches the assessment. The MonoTQ-InfoXLM-large model consistently exhibits superior Pearson correlation scores across a majority of the language pairs, accentuating its robust performance characteristics.

From our experiment results, as shown in Table 1, it's notable that the single-model configuration of MonoTQ-InfoXLM-large and MonoTQ-XLMV outperform ensemble-TQ for the majority of the language pairs. Observing the results outlined in both Table 1 and Table 2, it becomes evident that MonoTQ-InfoXLM-large and MonoTQ-XLMV not only outperform other systems among our own proposed approaches, they also exhibit a competitive performance with the best-performing system in the WMT23 sentence-level shared-task. MonoTQ-InfoXLM-large shows a very close Spearman correlation score with the winning system of the WMT23 sentence-level task for the En-Gu, En-Hi and En-Ta language pairs. Also, MonoTQ-XLMV shows the highest Spearman

No.	Name	DiskFootPrint (Bytes)
1	Unbabel-IST	42,868,104,221
2	IOL Research	2,357,242,105
3	HW-TSC	27,730,527,504
4	MMT	2,448,132,038
5	SurreyAI-ensembleTQ	7,945,689,496
6	SurreyAI-MonoTQ-XLMV	3,221,225,472
7	SurreyAI-MonoTQ-InfoXLM-large	2,362,232,012
8	SurreyAI-MonoTQ-XLMR-large	2,254,857,830

Table 3: Rows 1-5 display the disk footprint of ensemble model submissions related to the sentence-level task for WMT23. Meanwhile, Rows 6-8 present the disk footprint of our TQ models with single model configuration.

correlation score for the En-Te language pair. This observation raises the question that do the practice of ensembling always guarantees performance enhancement. Table 3 presents the memory requirements of both ensemble approaches and single-model configurations. Interestingly, in most cases ensemble models demand significantly more memory space than single-model setups, despite only offering a marginal boost in performance. This observation prompts us to reconsider the efficiency of employing ensemble models.

The conducted experiments across mid-resourced and low-resourced language pairs unravel intricate performance dynamics among various models.

7 Conclusion

This paper comprehensively evaluates the proposed architecture within the context of sentence-level direct quality assessment, employing diverse encoder-based pre-trained models. Our investigation notably highlights the enhanced performance attributed to the MonoTQ-InfoXLM-large, which surpasses the other configuration approaches, namely MonoTQ-XLMV, ensembleTQ strategy and MonoTQ-XLMR-large. While our outcomes in the WMT23 sentence-level Direct Assessment task did not attain peak performance, they nevertheless exhibited a marked improvement over the baseline and showed notable performance scores close to the winning systems.

Looking ahead, our research trajectory anticipates a continued exploration of quality estimation employing large language models. This involves further experimentation encompassing a broader spectrum of low-resourced language pairs. These forthcoming endeavours aspire to deepen our insights into the intricacies of direct quality assessment and contribute to advancing the frontiers of natural language processing. Also, we are focused on continuing the experimentation of pre-trained language models incorporated into different QE frameworks.

References

Keqin Bao, Yu Wan, Dayiheng Liu, Baosong Yang, Wenqiang Lei, Xiangnan He, Derek F. Wong, and Jun Xie. 2022. Alibaba-translate China's submission for WMT 2022 quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 597–605, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Xiang Geng, Yu Zhang, Shujian Huang, Shimin Tao, Hao Yang, and Jiajun Chen. 2022. NJUNLP's participation for the WMT2022 quality estimation

shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 615–620, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, António Góis, M. Amin Farajian, António V. Lopes, and André F. T. Martins. 2019a. Unbabel's participation in the WMT19 translation quality estimation shared task. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 78–84, Florence, Italy. Association for Computational Linguistics.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019b. OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122, Florence, Italy. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. Xlm-v: Overcoming the vocabulary bottleneck in multilingual masked language models. *ArXiv*, abs/2301.10472.

Seunghyun Lim and Jeonghyeok Park. 2022. Papago's submission to the WMT22 quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 627–633, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020a. TransQuest at WMT2020: Sentence-level direct assessment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1049–1055, Online. Association for Computational Linguistics.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020b. TransQuest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5070–5081, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2021. An exploratory analysis of multilingual word-level quality estimation with cross-lingual transformers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 434–440, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Carolina Scarton and Lucia Specia. 2014. Document-level translation quality estimation: exploring discourse and pseudo-references. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pages 101–108, Dubrovnik, Croatia. European Association for Machine Translation.

Philip Sedgwick. 2014. Spearman's rank correlation coefficient. *Bmj*, 349.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón F. Astudillo, and André F. T. Martins. 2018. Findings of the WMT 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 689–709, Belgium, Brussels. Association for Computational Linguistics.

Lucia Specia, Gustavo Paetzold, and Carolina Scarton. 2015. Multi-level translation quality prediction with QuEst++. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 115–120, Beijing, China. Association for Computational Linguistics and The Asian Federation of Natural Language Processing.

Hai Wang, Dian Yu, Kai Sun, Jianshu Chen, and Dong Yu. 2019. Improving pre-trained multilingual model with vocabulary expansion. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 316–327, Hong Kong, China. Association for Computational Linguistics.

Yige Xu, Xipeng Qiu, Ligao Zhou, and Xuanjing Huang. 2020. Improving bert fine-tuning via self-ensemble and self-distillation. *arXiv preprint arXiv:2002.10345*.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

MMT's Submission for the WMT 2023 Quality Estimation Shared Task

Yulong Wu[1], Viktor Schlegel[1, 2], Daniel Beck[3] and Riza Batista-Navarro[1]
[1] Department of Computer Science, University of Manchester, United Kingdom
[2] ASUS Intelligent Cloud Services (AICS), Singapore
[3] School of Computing and Information Systems, University of Melbourne, Australia
{yulong.wu, riza.batista}@manchester.ac.uk
viktor_schlegel@asus.com, d.beck@unimelb.edu.au

Abstract

This paper presents our submission to the WMT 2023 Quality Estimation (QE) shared task 1 (sentence-level subtask). We propose a straightforward training data augmentation approach aimed at improving the correlation between QE model predictions and human quality assessments. Utilising eleven data augmentation approaches and six distinct language pairs, we systematically create augmented training sets by individually applying each method to the original training set of each respective language pair. By evaluating the performance gap between the model before and after training on the augmented dataset, as measured on the development set, we assess the effectiveness of each augmentation method. Experimental results reveal that synonym replacement via the Paraphrase Database (PPDB) yields the most substantial performance boost for language pairs English-German, English-Marathi and English-Gujarati, while for the remaining language pairs, methods such as contextual word embeddings-based words insertion, back translation, and direct paraphrasing prove to be more effective. Training the model on a more diverse and larger set of samples does confer further performance improvements for certain language pairs, albeit to a marginal extent, and this phenomenon is not universally applicable. At the time of submission, we select the model trained on the augmented dataset constructed using the respective most effective method to generate predictions for the test set in each language pair, except for the English-German. Despite not being highly competitive, our system consistently surpasses the baseline performance on most language pairs and secures a third-place ranking in the English-Marathi[1].

1 Introduction

Quality Estimation (QE) strives to assess the output of Machine Translation (MT) systems without the availability of a reference translation of known high quality (Blatz et al., 2004; Specia et al., 2009, 2013; Kanojia et al., 2021). This capability serves as a valuable asset for expediting and cost-effectively facilitating the evaluation phases throughout the development cycle of MT systems.

In this paper, we describe our contribution to the QE shared task at the Eighth Conference on Machine Translation (WMT23). We participate in the Task 1 of the shared task and we specifically focus on the sentence-level subtask, which centers on predicting the quality score of neural MT outputs at the sentence level without access to reference translations. Our study encompasses six language pairs: *English-German (En-De)*, *English-Marathi (En-Mr)*, *English-Hindi (En-Hi)*, *English-Tamil (En-Ta)*, *English-Telegu (En-Te)*, *English-Gujarati (En-Gu)*, with annotations derived in two different ways: multi-dimensional quality metrics (MQM) (Freitag et al., 2021) and direct assessments (DA) (Fomicheva et al., 2022). Participating systems are assigned the task of predicting the quality score (MQM or DA) for each source-target sentence pair, and their performance is evaluated using Spearman's rank correlation coefficient as the primary metric, supplemented by the Kendall and Pearson coefficients as secondary metrics for assessment.

Our approach investigates the potential to enhance the performance of QE models by exposing them to a diverse range of training examples. To this end, we identify eleven different data augmentation methods and apply each of them individually to augment the training set for each language pair. Our results reveal that, for most language pairs, these methods result in varying degrees of performance improvement, with the most effective methods being synonym substitution using the PPDB, words insertion guided by contextual word embeddings, back-translation, and direct paraphrasing. We also show that for some language pairs, it

[1]Our code and data are available at https://github.com/Yulong-W/DataAug-QE.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 856–862
December 6–7, 2023.

is feasible to further enhance the model's performance by training it on an augmented set formed through the amalgamation of part or all of the said augmentation methods; however, the extent of improvement remains constrained. For each language pair except English-German, we generate predictions utilising the model trained on the augmented dataset constructed through the respective most effective method. Although our submission may not be considered highly competitive, they consistently achieve significantly improved performance compared to the organisers' baseline for the majority of language pairs. Notably, for the English-Marathi pair, our submission ranks third place with the Spearman score of 0.650. This observation indicates that the training data augmentation approach may hold particular promise and offer advantages when applied to the English-Marathi language pair.

2 Methodology

As mentioned above, we identified a total of eleven distinct data augmentation methods, as detailed in Table 1. For all the given source sentences and their corresponding MT hypothesis in the training dataset for each language pair, each method is independently applied only to the source sentences, leading to the creation of the respective transformed source-target sentence pairs. Our hypothesis posits that training the QE model on the augmented training set, which incorporates these transformed instances, holds the potential to bolster its performance. For each original instance, we generated one augmented sample per method and assigned to the augmented data the same quality score as the original translation hypothesis. However, it is noteworthy that certain methods, such as AS and RD, possess the potential to alter the meaning of the source-side sentence (Kanojia et al., 2021), consequently inducing changes in MT output and, by extension, the assigned quality label. In such instances, there is a likelihood of introducing noises to the augmented training dataset. A systematic exploration of the meaning-preserving capacity of these perturbation methods and the impact of those introduced noises at training time on the performance of the model necessitates further investigation.

3 Experiments

In this section, we describe our experimental settings, present the results achieved on the development and test sets, and perform analysis derived from our experimental findings. We additionally offer insights into the influence of the quantity of augmented training examples on the performance of the QE model.

	Data Augmentation Method
m_1	**WordNet-based Synonym Substitution (WSS):** Substitute words by WordNet's synonym (Fellbaum, 1998)
m_2	**PPDB-based Synonym Substitution (PSS):** Substitute words with synonyms from English PPDB (Pavlick et al., 2015)
m_3	**Antonym Substitution (AS):** Substitute random words with their antonyms
m_4	**Random Swap (RS):** Swap words in the sentence randomly
m_5	**Random Deletion (RD):** Delete words in the sentence randomly
m_6	**Spelling Mistake Substitution (SMS):** Substitute content words randomly by spelling mistake words dictionary
m_7	**GloVe Similarity-based Substitution (GSS):** Substitute words based on GloVe similarity (Pennington et al., 2014)
m_8	**Contextual Words Insertion (CWI):** Insert words using contextual word embeddings from the RoBERTa-base
m_9	**Contextual Words Substitution (CWS):** Substitute words by contextual word embeddings from the RoBERTa-base (Liu et al., 2019)
m_{10}	**Back Translation (BT)**
m_{11}	**Direct Paraphrasing (DP)**

Table 1: Various data augmentation methods.

3.1 Experimental Settings

Language Pairs (LPs). We conducted experiments on six language pairs. The training, development, and test datasets for each language pair utilised in our study are accessible via the shared task website[2], and we present the dataset statistics in Table 2. We applied each data augmentation method on the source sentences in the training set of each language pair.

Models. Our training methodology adheres to the PyTorch-based COMET framework (Rei et al., 2020), with the foundational pre-trained model be-

[2] https://wmt-qe-task.github.io/subtasks/task1/

LPs	Training	Development	Test
		MQM	
En-De	28909	1005	1897
		DA	
En-Mr	26000	1000	1086
En-Hi	7000	1000	1074
En-Ta	7000	1000	1075
En-Te	7000	1000	1075
En-Gu	7000	1000	1075

Table 2: Number of examples in the training, development and test set, respectively, for each language pair.

ing XLM-RoBERTa-large (Conneau et al., 2020). We fine-tuned the pre-trained XLM-RoBERTa-large model on the original and the augmented training sets for each language pair, respectively and evaluated them on the development set. The best-performing model was chosen from those trained on the corresponding augmented training datasets (in the case of English-German, the chosen model was trained on the augmented dataset created by applying the top four[3] effective data augmentation techniques to each source sentence) to generate the predictions on the test set. All experiments were conducted using 2 16GB Nvidia v100 GPUs.

Data Augmentation Methods. Methods WSS to BT: We utilised the NLPAug library (Ma, 2019) to perform the augmentation. In method PSS, we used the small size English PPDB (Pavlick et al., 2015). In the absence of any synonymous expressions documented for all the words within a source-side sentence in methods WSS and PSS, the augmented sample will persist unaltered in comparison to its original version. For methods WSS to CWS, the percentage of word will be augmented is set to the default value of 0.3, as in the implementation of the NLPAug library (Ma, 2019). In method BT, a sentence is translated from English to German, then back to English to obtain its paraphrased version (Ng et al., 2019). Method DP: Direct paraphrasing was performed by soliciting a Generative Pre-trained Transformer (GPT) (Brown et al., 2020) series model, specifically GPT-3.5-turbo, to generate responses for the prompt: *Generate a similar paraphrase for this sentence: [source sentence]*, using the OpenAI ChatGPT API.

[3]At the time of results submission, this number (i.e., 4) was randomly set. However, as illustrated in Figure 1, augmenting the training dataset for the English-German language pair using the best two methods yielded the most optimal performance.

3.2 Evaluation Results and Discussion

Table 3 illustrates the performance gap of the QE model on the development set before and after training on each augmented dataset created through the respective data augmentation method for each examined language pair. As can be seen from Table 3, in the majority of instances, the training data augmentation approaches demonstrated their effectiveness in enhancing the performance of the QE model. In the following, we discuss the observations for all the studied language pairs.

English-German. Method PSS exhibited the most significant performance improvement across all three evaluation metrics. Augmenting the training set with method CWI yielded the same improvements in terms of Spearman and Kendall correlations compared to augmenting it with PSS, albeit resulted a lower Pearson score. However, it was observed that presenting the model with modified training examples generated using method WSS and RS did not contribute to the enhancement of Spearman correlation. In fact, it even had an adverse effect, causing a slight reduction (0.3%) in the Kendall score.

English-{Marathi, Gujarati}. Training the model on the augmented set incorporating examples generated by substituting words with synonyms from PPDB (method PSS) proved to be the most effective approach in enhancing the correlation between the predictions of the model and human judgments of quality, with Spearman correlation increased by 6.8% and 7.1%, respectively. Other types of approaches also resulted in varying degrees of performance improvement.

English-{Hindi, Tamil}. For the English-Hindi language pair, augmenting the training set with both CWI and DP has been observed to yield identical improvements in terms of Spearman and Kendall correlations, emerging as the most effective approach. In the case of English-Tamil, the most notable enhancement was achieved by paraphrasing the source sentences in the original training dataset using the GPT-3.5-turbo model (method DP), as measured by Spearman and Kendall correlations. However, concerning the Pearson metric, method BT (back-translation) led to the most substantial improvements for both language pairs, amounting to 12.6% and 10.8%, respectively.

Method	En-De	En-Mr	En-Hi	En-Ta	En-Te	En-Gu	Average
				Spearman/Kendall/Pearson			
orig.	0.433/0.328/0.393	0.499/0.349/0.593	0.479/0.336/0.476	0.541/0.379/0.604	0.449/0.302/0.365	0.524/0.373/0.523	
WSS	0.433/0.327/0.404	0.518/0.363/0.607	0.492/0.346/0.512	0.548/0.386/0.652	0.430/0.291/0.362	0.536/0.384/0.579	
	0.0/-0.3/+2.8	+3.8/+4.0/+2.4	+2.7/+3.0/+7.6	+1.3/+1.8/+7.9	-4.2/-3.6/-0.8	+2.3/+2.9/+10.7	1.0
PSS	0.451/0.342/0.438	0.533/0.376/0.624	0.501/0.352/0.522	0.558/0.395/0.659	0.435/0.293/0.367	0.561/0.401/0.596	
	+4.2/+4.3/+11.5	**+6.8/+7.7/+5.2**	+4.6/+4.8/+9.7	+3.1/+4.2/+9.1	-3.1/-3.0/+0.5	**+7.1/+7.5/+14.0**	3.8
AS	0.442/0.335/0.408	0.516/0.364/0.616	0.503/0.354/0.528	0.542/0.382/0.662	0.431/0.294/0.360	0.547/0.392/0.586	
	+2.1/+2.1/+3.8	+3.4/+4.3/+3.9	+5.0/+5.4/+10.9	+0.2/+0.8/+9.6	-4.0/-2.6/-1.4	+4.4/+5.1/+12.0	1.8
RS	0.433/0.327/0.400	0.517/0.364/0.617	0.496/0.349/0.527	0.549/0.388/0.654	0.430/0.293/0.365	0.550/0.394/0.588	
	0.0/-0.3/+1.8	+3.6/+4.3/+4.0	+3.5/+3.9/+10.7	+1.5/+2.4/+8.3	-4.2/-3.0/0.0	+5.0/+5.6/+12.4	1.6
RD	0.442/0.335/0.424	0.507/0.356/0.601	0.494/0.347/0.526	0.551/0.389/0.648	0.437/0.296/0.373	0.552/0.397/0.591	
	+2.1/+2.1/+7.9	+1.6/+2.0/+1.3	+3.1/+3.3/+10.5	+1.8/+2.6/+7.3	-2.7/-2.0/+2.2	+5.3/+6.4/+13.0	1.9
SMS	0.435/0.329/0.404	0.517/0.363/0.604	0.500/0.351/0.525	0.547/0.386/0.656	0.439/0.300/0.369	0.552/0.396/0.590	
	+0.5/+0.3/+2.8	+3.6/+4.0/+1.9	+4.4/+4.5/+10.3	+1.1/+1.8/+8.6	-2.2/-0.7/+1.1	+5.3/+6.2/+12.8	2.1
GSS	0.440/0.333/0.417	0.521/0.366/0.610	0.500/0.352/0.522	0.555/0.392/0.653	0.435/0.298/0.369	0.547/0.392/0.578	
	+1.6/+1.5/+6.1	+4.4/+4.9/+2.9	+4.4/+4.8/+9.7	+2.6/+3.4/+8.1	-3.1/-1.3/+1.1	+4.4/+5.1/+10.5	2.4
CWI	0.451/0.342/0.430	0.518/0.364/0.619	0.509/0.358/0.535	0.546/0.385/0.661	0.450/0.308/0.384	0.554/0.397/0.590	
	+4.2/+4.3/+9.4	+3.8/+4.3/+4.4	**+6.3/+6.5**/+12.4	+0.9/+1.6/+9.4	**+0.2/+2.0/+5.2**	+5.7/+6.4/+12.8	3.5
CWS	0.444/0.337/0.412	0.513/0.359/0.609	0.506/0.355/0.525	0.554/0.392/0.656	0.442/0.302/0.377	0.543/0.387/0.588	
	+2.5/+2.7/+4.8	+2.8/+2.9/+2.7	+5.6/+5.7/+10.3	+2.4/+3.4/+8.6	-1.6/0.0/+3.3	+3.6/+3.8/+12.4	2.6
BT	0.441/0.334/0.423	0.522/0.366/0.612	0.504/0.354/0.536	0.559/0.397/0.669	0.435/0.295/0.373	0.552/0.395/0.593	
	+1.8/+1.8/+7.6	+4.6/+4.9/+3.2	+5.2/+5.4/**+12.6**	+3.3/+4.7/**+10.8**	-3.1/-2.3/+2.2	+5.3/+5.9/+13.4	2.8
DP	0.440/0.333/0.418	0.514/0.361/0.601	0.509/0.358/0.534	0.568/0.400/0.607	0.439/0.296/0.360	0.539/0.384/0.562	
	+1.6/+1.5/+6.4	+3.0/+3.4/+1.3	**+6.3/+6.5**/+12.2	**+5.0/+5.5**/+0.5	-2.2/-2.0/-1.4	+2.9/+2.9/+7.5	2.8

Table 3: The performance (%) of the QE model trained on the original, and the augmented training sets generated through applying the data augmentation methods, when evaluated on the development set for the examined language pairs. Values shown in the shaded areas are changes (%) relative to the original performance of the model, with the rightmost column shows their averages in terms of Spearman correlation. We highlight the values that denote the most substantial performance improvements across the Spearman, Kendall, and Pearson metrics.

English-Telegu. Our experimental training data augmentation approach was found to be notably ineffective when applied to the language pair English-Telegu. As shown in Table 3, in regard to Spearman and Kendall correlations, only method CWI yielded slight performance improvements, while the other approaches predominantly resulted in a decrease in the performance of the model. Indeed, these alternative approaches led to varying degrees of performance decline, with the most significant decrease being 4.2% in Spearman and 3.6% in Kendall, respectively. This may be attributed to the heightened sensitivity of English to Telegu translation concerning modifications applied to the source sentences. Consequently, noises might be introduced during the process of augmenting the training set, thereby contributing to a decline in the performance of the QE model.

Overall, our investigation revealed that, for the examined language pairs, method PSS yielded a relative performance increase of 3.8% on average, establishing itself as the most effective, with the second-best being CWI (3.5%). Interestingly, both method BT and method DP, designed for paraphrasing purposes, exhibited an identical average performance improvement of 2.8%. Conversely, the average increase was only 1.0% for method WSS, despite sharing the same objective of synonym substitution with method PSS. This suggests that employing synonym substitution via the English PPDB confers greater benefits to enhancing the performance of the QE model compared to performing it via WordNet. Furthermore, potential meaning alternation methods, such as AS and RD (Kanojia et al., 2021), yielded a lower average enhancement compared to some meaning-preservation methods like BT and DP. However, additional experimental confirmation is requisite.

3.3 Official Test Results

Based on the insights derived from Table 3, we systematically selected the most efficacious approach to augment the training set for each language pair and trained the respective model. Subsequently, we utilised each resulting model to generate predictions on the corresponding test dataset. For English-German language pair, it was observed that the performance of the QE model (0.303 Spearman),

when trained on the augmented dataset generated by applying PSS, was inferior to the baseline score determined during our initial test phase. Therefore, we took the initiative to curate a new training set wherein four augmented examples were generated for each original sample, employing the top four data augmentation methods identified as correspondingly effective. We then employed the re-trained model to generate quality predictions for English-German pair. The performance of our submitted models is presented in Table 4[4].

LPs	Spearman	Kendall	Pearson
	MQM		
En-De	0.316	0.237	0.221
	DA		
En-Mr	0.650	0.466	0.663
En-Hi	0.494	0.345	0.570
En-Ta	0.547	0.384	0.531
En-Te	0.337	0.228	0.281
En-Gu	0.540	0.386	0.581

Table 4: Official results of our submission to the WMT sentence-level QE shared task 2023.

Our most promising results were observed in language pair English-Marathi, where our submission ranked third among the six participating teams. This highlights the effectiveness of the training data augmentation approach in improving the capability of the QE model to precisely predict the quality score of English-Marathi translation pairs in the absence of a reference. However, when considering English-German, despite training the model on an augmented dataset with larger and more diverse samples, its performance still falls below the baseline score (0.340 Spearman). This discrepancy suggests that data augmentation approach may not be as efficient in enhancing the QE performance for this specific language pair. Nevertheless, we observed that this performance (0.316 Spearman) remains slightly superior to that achieved with the training set containing fewer augmented samples (0.303 Spearman), which indicates that increasing the number of augmented training examples might contribute to enhancing the performance of the model, and we provide further elaboration in Section 3.4 below. In contrast, for the remaining four language pairs we investigated, the performance of our submitted models consistently outperformed the baseline score. Specifically, our submission demonstrated a notable enhancement over the baseline score in Spearman correlation for English-Hindi (+0.213), English-Telugu (+0.144), and English-Gujarati (+0.203), while the improvement for English-Tamil was comparatively less pronounced. Despite the above-baseline performance achieved, our submission is presently ranked fifth in these language pairs, signifying the necessity for additional investigation and refinement of our approach to attain elevated performance levels.

[4]A comparison of our results with the organiser's baseline and submissions from other participating teams is available at http://www2.statmt.org/wmt23/quality-estimation-task_results.html.

3.4 Impact of Training Example Quantity

Thus far, a singular augmented example has been generated corresponding to each defined augmentation method for every original training sample in our studied language pairs, with the exception of English-German. To examine the impact of the number of augmented samples on the performance of the QE model and to explore potential complementarity among these augmentation techniques, we trained the models for each language pair on augmented training sets of varying sizes, generated by employing the respective top N effective augmentation methods (where N ranges from 1 to 11), and then assessed their performance, as shown in Figure 1.

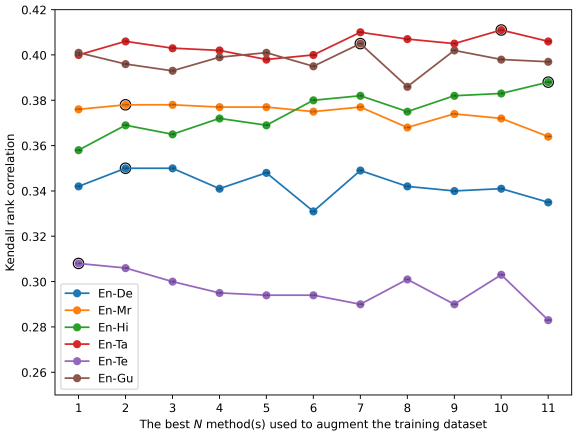


Figure 1: The performance of the QE models on the development set across six language pairs, trained on augmented datasets generated utilising the respective best N data augmentation methods. The optimal performance was denoted by encircling the respective data point with a black circle.

It can be seen from Figure 1 that for five lan-

guage pairs English-{Tamil, Gujarati, Marathi, Hindi, German}, increasing the number of augmented training samples can enhance the model's performance, although this phenomenon is not universal for certain language pairs, such as English-Tamil. However, we observed that there was a negligible degree of performance improvement across these five language pairs, with the most notable enhancement being merely 0.03 (from 0.358 to 0.388), as demonstrated in the case of English-Hindi. Even worse, for the language pair English-Telegu, exposing the model to a more diverse set of training examples resulted in a decline in performance. Notably, training the model on an augmented set comprising eleven augmented samples per original instance led to the nadir in performance, recording a value of 0.283. This underscores the constraints of current data augmentation methods in boosting the efficacy of the QE model, emphasizing the imperative to devise more effective approaches. Nonetheless, a positive insight has been discerned; the language pair English-Hindi appears to derive particular benefits from the augmentation of training examples. As the number of applied top N augmentation methods increased, the performance of the model consistently surpassed that of the model with only the best one applied, notwithstanding fluctuations in performance. Finally, based on the empirical findings depicted in Figure 1, definitive conclusion regarding the complementarity of specific data augmentation approaches cannot be drawn, as it is inherently specific to each language pair. For instance, the efficacy of combining the best two augmentation methods was observed in the English-{Marathi, German} pairs, whereas for English-{Tamil, Gujarati, Hindi}, optimal performance was attained through the amalgamation of the top 10, 7, and 11 training data augmentation methods, respectively.

4 Conclusion

In this paper, we proposed a training data augmentation approach to the WMT 2023 sentence-level QE shared task. We systematically identified eleven various data augmentation methods and applied each of them individually on the source-side sentences to generate augmented training samples for the six studied language pairs. The experimental results demonstrated that in most cases, these methods can enhance the correlation between the predictions of the QE model and human-provided quality scores to varying degrees, albeit not to a significant extent. In addition, we show that training the model on the augmented set, generated through the combination of these methods, contributed further to performance enhancement, although this phenomenon was not universally observed and the degree of improvement was at a negligible level. Our methodology yielded a third-ranking outcome for English-Marathi and a fifth-place ranking for other DA annotated language pairs, among the submissions from the six teams. In terms of future work, we intend to explore other more effective augmentation approaches and extend our study to encompass a more diverse array of language pairs and QE models.

Limitations

The work presented in this paper should be considered preliminary, given that we exclusively conducted experiments employing a training data augmentation approach and assessed its impact solely on the original development set. There is ample room for further exploration into the robustness of the QE model without any augmentation interventions on the studied perturbations and the impact of these proposed perturbations, when applied during training, on the capability of the QE systems to identify critical errors in translation resulting from modifications to the source sentences. Moreover, the extent to which the introduced perturbations may alter the meaning of the source-side sentences remains unclear, necessitating further investigation.

Acknowledgements

The authors would like to thank the University of Manchester Department of Computer Science Kilburn Scholarship, the Manchester-Melbourne-Toronto Research Fund 2022, and the Turing Scheme for supporting this work. We also express our sincere gratitude for the invaluable comments and suggestions provided by the anonymous reviewers and acknowledge the support of the Computational Shared Facility at The University of Manchester in facilitating the execution of our experiments.

References

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation

for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 315–321, Geneva, Switzerland. COLING.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2022. MLQE-PE: A multilingual quality estimation and post-editing dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4963–4974, Marseille, France. European Language Resources Association.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Diptesh Kanojia, Marina Fomicheva, Tharindu Ranasinghe, Frédéric Blain, Constantin Orăsan, and Lucia Specia. 2021. Pushing the right buttons: Adversarial evaluation of quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 625–638, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 425–430, Beijing, China. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Lucia Specia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn. 2013. QuEst - a translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 79–84, Sofia, Bulgaria. Association for Computational Linguistics.

Lucia Specia, Marco Turchi, Nicola Cancedda, Nello Cristianini, and Marc Dymetman. 2009. Estimating the sentence-level quality of machine translation systems. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, Barcelona, Spain. European Association for Machine Translation.

IOL Research's Submission for WMT 2023 Quality Estimation Shared Task

Zeyu Yan, Wenbo Zhang, Qiaobo Deng,
Hongbao Mao, Jie Cai, Zhengyu He
Transn IOL Research, Wuhan, China
{zeyu.yan,albert01.zhang,qiaobo.deng,hubben.mao,jay.cai,steven.he}@transn.com

Abstract

This paper presents the submissions of IOL Research in WMT 2023 quality estimation shared task. We participate in task 1 Quality Estimation on both sentence and word levels, which predicts sentence quality score and word quality tags. Our system is a cross-lingual and multitask model for both sentence and word levels. We utilize several multilingual Pretrained Language Models (PLMs) as backbones and build task modules on them to achieve better predictions. A regression module on PLM is used to predict sentence level score and word tagging layer is used to classify the tag of each word in the translation based on the encoded representations from PLM. Each PLM is pretrained on quality estimation and metrics data from the previous WMT tasks before finetuning on training data this year. Furthermore, we integrate predictions from different models for better performance while the weights of each model are automatically searched and optimized by performance on Dev set. Our method achieves competitive results.

1 Introduction

Quality Estimation (QE) is the task of predicting the quality of a target machine translation without using reference texts or human inputs (Specia et al., 2018). Since machine translation is in high demand nowadays, the development of QE system becomes crucial for the broad application of machine translation. In WMT 2023 Quality Estimation shared task, there are two tasks: quality estimation and fine-grained error span detection. This paper describes our submission to task 1 quality estimation in both sentence and word levels in detail.

Considering the powerful capability and widely used in previous QE tasks of pretrained language models (PLMs) (Zerva et al., 2022; Specia et al., 2021), our method utilizes different multilingual PLMs to encode *source-translation* sentence pairs and predict sentence-level scores or word-level tags. Such PLMs are pretrained on various languages which could show incredible ability when trained QE models are transferred to unseen language pairs. Meanwhile, extra task modules are added to PLMs to boost the interaction between source and translation sentences to make better predictions. Also, it is common that using other task data similar to QE can further improve the performance of QE. According to the results from previous years' QE tasks, we use the data from QE and Metrics tasks from previous years' WMT tasks, as well as Automatic Post-Editing (APE) data, to pretrain PLMs before training on data of this year.

Moreover, ensemble methods of different models are explored in sentence and word level tasks. For sentence level, we sum scores with weights from different models which are filtered by the performance on the Dev set. As for word level, we use voting or weighted sum of tag probabilities to get the final predicted tags. Taking zero-shot language pairs into account, we choose the best model evaluated on other language pairs to test if they can generalize to unseen language pairs.

2 Quality Estimation Task

2.1 Task description

WMT 2023 Quality Estimation task 1 contains two tasks. The sentence level task aims at predicting a quality score for *translation* and the word level task is to classify a quality tag for each word in *translation*. Both tasks have zero-shot language pairs to test the generalization ability of QE models and use the same *source-translation* pairs for each language pair.

Sentence level There are two types of quality scores. One is the Direct Assessments (DA) score which is given by human annotators for each *source-translation* pair. The other is the Multi-dimensional Quality Metrics (MQM) score which is defined and computed under MQM methodology.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 863–871
December 6–7, 2023.

train_sent	train_word	train_mtl
224195	263184	105992

Table 1: The statistics of train data

	Dev	Test
En-De	511	1897
Zh-En	505	1677
En-Mr	1000	1086
En-Gu	1000	1075
En-Hi	1000	1074
En-Ta	1000	1075
En-Te	1000	1075
He-En	-	1182
En-Fa	-	1000

Table 2: The statistics of dev and test data

A regression model is always employed to predict quality scores.

Word level The tags of words in *translation* are annotated by human annotators according to the MQM or DA annotations. This task requires predicting an OK or BAD for each word in *translation* given *source-translation* pairs. HTER (Specia and Farzindar, 2010)-like scores for translations can be collected by calculating the ratio of 'BAD' tags in tag sequence of *translations*. For example, given a tag sequence "OK OK BAD BAD OK", an HTER-like score is deduced by computing 2/5=0.4.

Data QE task provides official train and dev datasets gathered from competitions of previous years and the statistics are shown in Table 1 and Table 2. On account of the task similarity to QE, we also collect the MQM data (Freitag et al., 2021a,b) from previous WMT Metrics tasks[1] and APE data from QT21 (Specia et al., 2017) and APE-QUEST (Depraetere et al., 2020) to do further pretraining. We calculate HTER-like score for each *source-translation* pair in APE data for the purpose of merging with those of DA and MQM.

3 Method

3.1 Model architecture

We design distinct task modules on top of encoders for regression on sentence level and sequence tagging on word level. Source and translation texts are concatenated and input into the encoder and then task modules to get scores or tags. Our model architecture is illustrated in Fig.1.

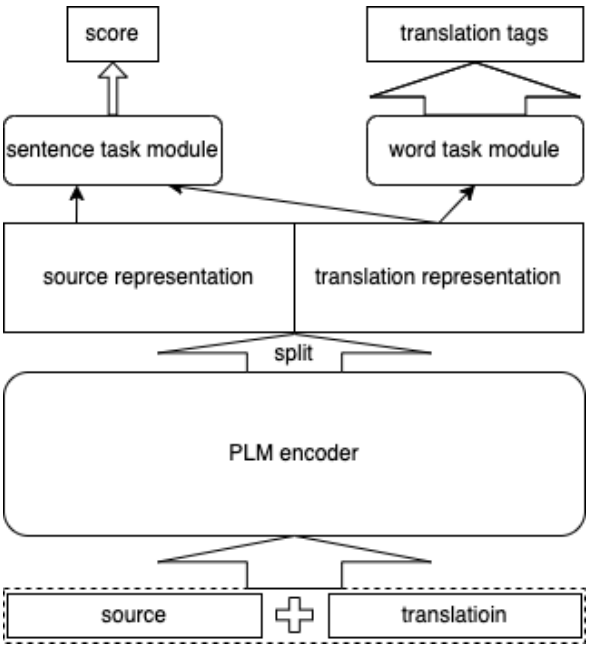


Figure 1: Model architecture with task modules for sentence-level scoring and word-level tagging

Sentence regression module Inspired by ESIM (Chen et al., 2017) and RE2 (Yang et al., 2019), the cross attention between *source* and *translation* reflects the similarity between words in different languages. Also considering that different layers in a transformer (Vaswani et al., 2017) based PLM catch different granularities of features of *source* and *translation* (Jawahar et al., 2019), we determine to combine these two kinds of methods to strengthen the representations of *source* and *translation*. In detail, for *source* s and *translation* t respectively, mixed layer-wise representations s_{mix} and t_{mix} from a PLM with L layers are computed in Eq. 1~Eq. 4.

$$s^l = mean\_pooling([s_1^l, s_2^l, ..., s_m^l]), \quad (1)$$

$$t^l = mean\_pooling([t_1^l, t_2^l, ..., t_n^l]), \quad (2)$$

$$s_{mix} = \sum_{l=1}^{L} w_s^l * s^l, \; where \sum_{l=1}^{L} w_s^l = 1 \quad (3)$$

$$t_{mix} = \sum_{l=1}^{L} w_t^l * t^l, \; where \sum_{l=1}^{L} w_t^l = 1 \quad (4)$$

Then cross attention outputs s_{ca} and t_{ca} from the last layer of PLM are calculated to get token level

[1] https://github.com/Unbabel/COMET

interactions between *source* and *translation* as shown in Eq. 5 ~Eq. 9.

$$e_{ij} = s_i^T t_j \tag{5}$$

$$s_i^{ca} = \sum_{j=1}^{n} \frac{exp(e_{ij})}{\sum_{k=1}^{n} exp(e_{ik})} t_j, \ \forall i \in [1, 2, ..., m] \tag{6}$$

$$t_j^{ca} = \sum_{i=1}^{m} \frac{exp(e_{ij})}{\sum_{k=1}^{m} exp(e_{kj})} s_i, \ \forall j \in [1, 2, ..., n] \tag{7}$$

$$s_{ca} = mean\_pooling([s_1^{ca}, s_2^{ca}, ..., s_m^{ca}]) \tag{8}$$

$$t_{ca} = mean\_pooling([t_1^{ca}, t_2^{ca}, ..., t_n^{ca}]) \tag{9}$$

Next, features of *source* and *translation* are fused separately to transform into a combined representation through feedforward network (FFN) layer by Eq. 10 and Eq. 11.

$$s_{comb} = FFN([s_{ca}; s_{mix}; |s_{ca} - s_{mix}|; s_{ca} * s_{mix}]) \tag{10}$$

$$t_{comb} = FFN([t_{ca}; t_{mix}; |t_{ca} - t_{mix}|; t_{ca} * t_{mix}]) \tag{11}$$

Finally, the sentence-level score is obtained by another FFN layer in Eq. 12.

$$score = FFN([s_{comb}; t_{comb}]) \tag{12}$$

Word tagging module We choose two distinct modules to generate tags for words after encoded by PLM. A Bidirectional-LSTM (Hochreiter and Schmidhuber, 1997) (BiLSTM) layer is added to enhance the interaction between the representations of *source* and *translation*, and a FFN layer on it to predict tags in *translation*. Another kind of module only adopts a FFN layer to generate tag predictions to avoid overfitting on training data.

Multitask combination In order to boost the individual performance of sentence level and word level models, we propose a multitask training approach. Both the regression module and tagging module are added to the encoder which predicts the sentence score like DA or MQM and word tags simultaneously. For language pairs that have no DA or MQM data but only word tags, we take HTER scores as sentence scores. We train word-level models by optimizing the prediction of HTER scores and word tags simultaneously. To not damage the potentiality of tagging module, some simple regression modules are used when doing multitask training, including

$$score = FFN(mean\_pooling(\mathbf{t}_{[1:n]})) \tag{13}$$

and

$$score = FFN([\bar{s}; \bar{t}; |\bar{s} - \bar{t}|; \bar{s} * \bar{t}]) \tag{14}$$

where $\mathbf{t}_{[1:n]}$ is the list of word representations of *translation* from tagging module, and $\bar{s}$ and $\bar{t}$ are the mean representations of words' representations of *source* and *translation* from the encoder.

Loss The losses for score regression, word tagging and multitask training are described as follows:

$$\mathcal{L}_{sent} = (score_{pred} - score_{true})^2 \tag{15}$$

$$\mathcal{L}_{word} = -\frac{1}{n} \sum_{i=1}^{n} \log p(y_i) \tag{16}$$

$$\mathcal{L}_{multitask} = \mathcal{L}_{sent} + \mathcal{L}_{word} \tag{17}$$

where $p(y_i)$ is the probability of OK/BAD tag from the model.

3.2 Score refinement

According to the similar definitions and score intervals of DA and MQM, we transform the score s out of [-1, 1] as close to [-1,1] as possible while keeping the Spearman correlation coefficient unchanged using Eq. 18 to lessen the need for predicting extreme values during training.

$$s' = \begin{cases} \text{(s+1)*0.1-1,} & \text{s} < -1 \\ \text{s,} & \text{-1} \leq \text{s} \leq 1 \\ \text{(s-1)*0.1+1,} & \text{s} > 1 \end{cases} \tag{18}$$

3.3 Encoder selection

QE requires texts from different languages as input, so we take multilingual PLMs as encoders which are pretrained on colossal multilingual corpus. The following PLMs are selected as encoders: XLM-Roberta-Large (Conneau et al., 2020)[2], RemBert (Chung et al., 2021)[3], InfoXLM-Large (Chi et al., 2021)[4] and mDeBERTa (He et al., 2021)[5]. Each PLM is combined with different task modules for training.

[2]https://huggingface.co/xlm-roberta-large
[3]https://huggingface.co/google/rembert
[4]https://huggingface.co/microsoft/infoxlm-large
[5]https://huggingface.co/microsoft/mdeberta-v3-base

3.4 Model Training

We first pretrain encoders with a simple regression head to do regression on WMT Metrics and HTER data while retaining the checkpoints of encoders with the best performance on Dev set. When using WMT Metrics data, we train two versions of models where one uses DA data only and the other uses a mix of DA and MQM data. Subsequently, we drop the regression head and then finetune the pretrained encoder with different task modules on multilingual QE data. In order to eliminate the possible side effect of position variation in *translation*, we swap the input order of *source* and *translation* as a comparison. We conduct single-task and multitask training for both sentence and word levels.

3.5 Ensemble methods

Sentence level For each language pair having training data, we randomly search weights for the weighted sum of the top 10 models in accordance with the Spearman correlation coefficient on Dev set. As for zero-shot language pairs, we pick the best two or three trained models from those language pairs having training data individually then predict and average the scores from them.

Word level We propose three strategies of tag prediction ensemble for each word. At first, for each language pair having training data, the top 10 models with the best Matthews Correlation Coefficient on Dev set are picked out. Therefore each word in *translation* has 10 predicted tags or 10 probability pairs of (OK, BAD) from different models. The final tag of one word is acquired in one of three ways:

1 if one of 10 tags is BAD, the final tag is BAD;

2 if one of 10 tags is OK, the final tag is OK;

3 if the weighted sum of probabilities of OK is larger than that of BAD, the final tag is OK, and vice versa.

When utilizing the third one, the weights of models are searched randomly as in sentence-level ensemble. As for zero-shot language pairs, we pick the best two trained models from those language pairs having training data individually and apply one of the above strategies to get final predictions.

4 Experiments

4.1 Settings

All our models are completed with PyTorch and transformers (Wolf et al., 2020)[6] and trained on NVIDIA GeForce RTX 3090 24G for the pretraining and finetuning described in 3.4. Models are trained with AdamW (Loshchilov and Hutter, 2017) with learning rate of 1e-5, max sequence length of 230, batch size of 16 and 3 epochs. Models with different task modules are optimized by selecting the checkpoint with the best Spearman correlation coefficient or Matthews Correlation Coefficient (MCC) on Dev set for each language pair separately. Three versions of PLM are pretrained as described in 3.4 for each combination of language pair and PLM, which are listed in the order of "DA-only, DA+MQM, HTER" in Table 4, Table 7 and Table 8 while Table 3 are only "DA-only, DA+MQM" for each language pair. Optuna (Akiba et al., 2019) is used to search the weights of model ensembles described in 3.5.

4.2 Results and Analysis

Sentence level For results in Table 3 of Dev set with MQM annotations, results based on mDeBERTa perform best in all settings. Models with PLMs pretrained on "DA-only" data achieve better results than those "DA+MQM" models which indicates that the difference in score range between DA and MQM has a great effect. For Table 7 of Dev set with DA annotations, models pretrained on "DA-only" data perform best among different combinations of PLMs and language pairs. Also, InfoXLM and XLM-Roberta-Large show higher correlations than other PLMs. Meanwhile, the score refinement defined in 3.2 has a positive impact in both Table 3 and Table 7 which suggests the necessity to unify the range of different scores. However, correlations of different PLMs vary a lot for each language pair which suggests we still have room for improvement. Also, when using multitask training, the Spearman correlation coefficient increases compared to only training on sentence-level data. The "DA+MQM" data improves the performance of En-De while becoming worse on Zh-En.

Word level The results in Table 4 indicate that pretraining data, PLM and task modules affect the model performance to varying degrees. Since HTER data is most related to word-level task, the

[6]https://github.com/huggingface/transformers

sentence level	MQM			
	En-De		Zh-En	
	DA-only			
XLM-Roberta-Large	0.5162	0.4219	0.3424	0.3028
mDeBERTa	0.5467	0.5281	0.3310	0.3717
RemBert	0.5040	0.4231	0.3048	0.2948
InfoXLM	0.5295	0.3786	0.3670	0.2881
	DA+MQM			
XLM-Roberta-Large	0.5346	0.4459	0.2889	0.2495
mDeBERTa	0.5668	0.5470	0.3110	0.3597
RemBert	0.5141	0.4323	0.3042	0.2822
InfoXLM	0.5342	0.4451	0.3039	0.2793
	DA-only w/ score_refine			
XLM-Roberta-Large	0.5218	0.4277	0.3254	0.2772
mDeBERTa	0.5435	0.5202	0.3319	0.3594
RemBert	0.5144	0.4092	0.2894	0.3080
InfoXLM	0.5266	0.4047	0.3734	0.2945
	DA+MQM w/ score_refine			
XLM-Roberta-Large	0.5386	0.4561	0.3005	0.2473
mDeBERTa	0.5728	0.5494	0.3227	0.3547
RemBert	0.5092	0.4302	0.2973	0.2840
InfoXLM	0.5309	0.4599	0.3037	0.2863

Table 3: Spearman correlation on Dev of sentence level on combinations of training data and score refinement(optional)

results based on pretraining on HTER data are best. Besides, models with RemBert or InfoXLM on En-De give bad results while models with BiLSTM as task module on Zh-En overfit on Dev set when submitting to test. In addition, swapping the order of *source* and *translation* has no improvement. For En-De and En-Mr, training on word-level data only is better than multitask training.

Multitask As shown in Table 8, multitask training improves the correlation of sentence-level task on all language pairs while only MCC of Zh-En grows. The score refinement method raises the correlation of word-level task obviously compared to models without applying score refinement. Yet, it does not always have a positive effect on sentence-level task. The multitask training for Zh-En avoids overfitting on Dev set and using BiLSTM as task module surpasses using FFN. Different PLMs will perform better if combined with specific task modules, which needs further experiments.

Ensemble The official results of models ensemble on dev and test for sentence level and word level are shown in Table 5 and Table 6 respectively. The ensemble method outperforms single model performance by a large margin. Our models have competitive results on all language pairs.

5 Conclusion

This paper describes our work for WMT 2023 Quality Estimation Task 1 on both sentence level and word level. With the help of PLMs and extra data, we can train better representations of *source* text and its *translation* for quality estimation task. We also experiment with diverse combinations of PLMs, task modules, and pretraining datasets. We find that QE systems for certain language pairs need to adopt particular combinations to acquire improvement, which reveals that there are distinct characteristics between languages. Such features make it hard to build one model for all languages, especially those without labeled data. The multitask training approach shows obvious improvements and prevents models from overfitting. Besides, the score refinement trick does not always give us positive feedback which suggests the number range is not the only factor to train on DA and MQM data properly. As expected, the ensemble method makes the predictions have a higher correlation with the ground truth. For future work, we will explore more profitable pretraining techniques for quality estimation and efficient modules that work well for various language pairs.

Limitations

Although our method has shown competitive results on most language pairs, evaluation results on zero-shot language pairs suggest that the model is not so powerful in generalization and relies on manual adjustment to some extent like choosing the weights among different models in the ensemble. Such operations could affect the model performance when transferring to unseen language pairs. Furthermore, we only designed two kinds of modules to generate tags in word-level task with slight improvement over baselines. It will be a potential research area to design more efficient prediction modules that can predict more accurate tags and we leave it as future work.

Also, other training configurations like weight decay and layer-wise learning rate decay were not experimented with sufficiently. Due to the discrepancy between training loss and evaluation metric, the choice of loss was a critical factor in model performance which was unexplored. Lastly, the limited amount of data constrained the improvement of models and overfitting on Dev set still has a great effect on optimization. We hope these analyses can promote the research of quality estimation.

word level	En-De			Zh-En			En-Mr		
BiLSTM + regression(Eq. 13)									
mDeBERTa	0.3354	0.3364	0.3388	0.4483	0.4868	0.4447	0.3443	0.3500	0.3566
RemBert	0.0370	0.0160	0.0076	0.4842	0.4140	0.4736	0.3657	0.3601	0.3637
InfoXLM	0.0327	0.0456	0.0288	0.5491	0.4656	0.5249	0.3466	0.3385	0.3603
FFN + regression(Eq. 14)									
mDeBERTa	0.3206	0.3306	0.3315	0.4666	0.5013	0.4727	0.3399	0.3396	0.3443
RemBert	0.3213	0.2477	0.3313	0.4715	0.4993	0.4575	0.3724	0.3158	0.3504
InfoXLM	0.2972	0.2905	0.3042	0.5411	0.4860	0.5230	0.3554	0.3407	0.3601
FFN + regression(Eq. 14) w/ swap_order									
mDeBERTa	0.3167	0.3451	0.3306	0.5252	0.4951	0.4506	0.3305	0.3339	0.3469
RemBert	0.3123	0.2752	0.3023	0.4547	0.4513	0.4715	0.3610	0.3448	0.3153
InfoXLM	0.2969	0.2851	0.2957	0.5167	0.5032	0.5549	0.3520	0.3277	0.3626

Table 4: Spearman correlation on Dev of word level on combinations of tagging modules(BiLSTM/FFN) and regression modules with swapping orders(optional)

	Dev	Test
En-De	0.612	0.483
Zh-En	0.403	0.482
En-Mr	0.626	0.505
En-Gu	0.706	0.695
En-Hi	0.603	0.600
En-Ta	0.708	0.740
En-Te	0.474	0.376
He-En	-	0.575
Multilingual	-	0.513

Table 5: Spearman correlation of sentence level on Dev and Test

	Dev	Test
En-De	0.343	0.256
Zh-En	0.221	0.250
En-Mr	0.398	0.334
He-En	-	0.359
En-Fa	-	0.351
Multilingual	-	0.298

Table 6: MCC of word level on Dev and Test

Ethics Statement

This work follows all the rules of ACL Ethics Policy during the experiments of training and evaluation. The data used in this work are publicly available and widely used or provided by the organization of the competition. And to the best of the authors' knowledge, we do not foresee any risks against the ACL Ethics Policy.

Acknowledgements

The participants would like to express heartfelt thanks to the committee and the organizers of the WMT Quality Estimation Shared Task. We would also like to show our gratitude to the reviewers for their invaluable suggestions. This work is supported by Transn IOL Technology Co., Ltd.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An

information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Heidi Depraetere, Joachim Van den Bogaert, Sara Szoc, and Tom Vanallemeersch. 2020. APE-QUEST: an MT quality gate. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 473–474, Lisboa, Portugal. European Association for Machine Translation.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Chrysoula Zerva, Zhenhao Li, Vishrav Chaudhary, and André F. T. Martins. 2021. Findings of the WMT 2021 shared task on quality estimation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 684–725, Online. Association for Computational Linguistics.

Lucia Specia and Atefeh Farzindar. 2010. Estimating machine translation post-editing effort with HTER. In *Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry*, pages 33–43, Denver, Colorado, USA. Association for Machine Translation in the Americas.

Lucia Specia, Kim Harris, Frédéric Blain, Aljoscha Burchardt, Viviven Macketanz, Inguna Skadiņa, Matteo Negri, , and Marco Turchi. 2017. Translation quality and productivity: A study on rich morphology languages. In *Proceedings of Machine Translation Summit XVI*, pages 55–71, Nagoya, Japan.

Lucia Specia, Carolina Scarton, and Gustavo Henrique Paetzold. 2018. Quality estimation for machine translation. *Synthesis Lectures on Human Language Technologies*, 11(1):1–162.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 5998–6008. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Runqi Yang, Jianhai Zhang, Xing Gao, Feng Ji, and Haiqing Chen. 2019. Simple and effective text matching with richer alignment features. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4699–4709, Florence, Italy. Association for Computational Linguistics.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

sentence level	Direct Assessment														
	En-Mr			En-Gu			En-Hi			En-Ta			En-Te		
DA-only															
XLM-Roberta-Large	0.5867	0.5725	0.5814	0.6554	0.6367	0.6472	0.5152	0.5082	0.5207	0.6672	0.6658	0.6580	0.4092	0.4198	0.3972
mDeBERTa	0.5628	0.5467	0.5554	0.6370	0.6058	0.6288	0.5083	0.5004	0.5237	0.6423	0.6119	0.6538	0.3405	0.3263	0.3477
RemBert	0.5556	0.5593	0.5714	0.6649	0.6295	0.6422	0.5547	0.5350	0.5470	0.6598	0.6439	0.6592	0.4312	0.4028	0.3990
InfoXLM	0.5658	0.5697	0.5693	0.6715	0.6569	0.6693	0.5317	0.5386	0.5266	0.6701	0.6650	0.6708	0.4211	0.4131	0.3852
DA+MQM															
XLM-Roberta-Large	0.5869	0.5761	0.5582	0.6671	0.6352	0.6354	0.5282	0.5125	0.5268	0.6561	0.6694	0.6692	0.4202	0.4343	0.4293
mDeBERTa	0.5683	0.5591	0.5588	0.6275	0.5972	0.6252	0.5101	0.5043	0.5128	0.6307	0.6183	0.6244	0.3529	0.3251	0.3260
RemBert	0.5558	0.5802	0.5583	0.6357	0.6215	0.6541	0.5484	0.5401	0.5643	0.6507	0.6513	0.6657	0.4050	0.4094	0.4083
InfoXLM	0.5691	0.5713	0.5691	0.6631	0.6460	0.6609	0.5246	0.5227	0.5039	0.6792	0.6659	0.6685	0.4394	0.3978	0.4189
DA-only w/ score_refine															
XLM-Roberta-Large	0.5810	0.5740	0.5803	0.6738	0.6389	0.6614	0.5273	0.5307	0.5338	0.6730	0.6651	0.6764	0.4398	0.4174	0.3924
mDeBERTa	0.5599	0.5675	0.5539	0.6335	0.6187	0.6216	0.5022	0.5167	0.5245	0.6639	0.6318	0.6525	0.3476	0.3250	0.3449
RemBert	0.5888	0.5442	0.5688	0.6651	0.6378	0.6596	0.5826	0.5252	0.5813	0.6685	0.6554	0.6643	0.4166	0.3896	0.4348
InfoXLM	0.5705	0.5571	0.5814	0.6806	0.6678	0.6768	0.5485	0.5410	0.5417	0.6847	0.6704	0.6814	0.4256	0.4150	0.4086
DA+MQM w/ score_refine															
XLM-Roberta-Large	0.5740	0.5555	0.5773	0.6568	0.6308	0.6466	0.5205	0.4811	0.5065	0.6726	0.6506	0.6566	0.4256	0.4281	0.4176
mDeBERTa	0.5523	0.5756	0.5630	0.6434	0.6266	0.6236	0.5138	0.5170	0.5044	0.6412	0.6264	0.6330	0.3343	0.3447	0.3562
RemBert	0.5744	0.5620	0.5724	0.6387	0.6339	0.6328	0.5734	0.4939	0.5656	0.6433	0.6648	0.6618	0.4156	0.4161	0.4076
InfoXLM	0.5746	0.5724	0.5804	0.6746	0.6522	0.6627	0.5308	0.5367	0.5144	0.6894	0.6655	0.6634	0.4299	0.422	0.4488

Table 7: Spearman correlation on Dev of sentence level on combinations of training data and score refinement(optional)

multitask	En-De			Zh-En			En-Mr		
				SRM + BiLSTM					
mDeBERTa	0.5724/0.2933	0.5360/0.3503	0.5749/0.3012	0.3063/0.2360	0.3383/0.2062	0.3170/0.2412	0.5783/0.2294	0.5808/0.2381	0.5707/0.2086
RemBert	0.5390/0.0018	0.4353/0.0013	0.5340/0.0052	0.2882/0.1920	0.2661/0.1804	0.3108/0.1879	0.5895/0.2934	0.5710/0.2515	0.5926/0.3205
InfoXLM	0.5206/-0.0004	0.4342/0.0377	0.5033/0.0087	0.3237/0.2826	0.2791/0.2230	0.2980/0.2791	0.5712/0.2860	0.5739/0.2788	0.5801/0.3151
				SRM + FFN					
mDeBERTa	0.5638/0.3036	0.5389/0.3070	0.5738/0.3055	0.3113/0.2212	0.3377/0.2091	0.3140/0.2187	0.5859/0.2581	0.5669/0.2555	0.5829/0.2942
RemBert	0.5439/0.2475	0.4288/0.2509	0.5134/0.2869	0.2834/0.1774	0.2878/0.1558	0.3069/0.1516	0.5787/0.2982	0.5674/0.2615	0.5911/0.3036
InfoXLM	0.5317/0.3161	0.4254/0.1994	0.5358/0.2149	0.3487/0.3283	0.2721/0.2043	0.3114/0.2935	0.5703/0.3018	0.5668/0.2896	0.5799/0.3169
				SRM + BiLSTM w/ score_refine					
mDeBERTa	0.5580/0.3028	0.5411/0.3461	0.5597/0.2766	0.3281/0.2308	0.3438/0.2116	0.3143/0.2027	0.5877/0.2796	0.5709/0.2959	0.5834/0.2819
RemBert	0.5457/0.0080	0.4425/0.0119	0.5239/0.0051	0.2750/0.1871	0.2968/0.1695	0.3332/0.1684	0.5762/0.3176	0.5962/0.3233	0.5868/0.3250
InfoXLM	0.5272/0.0088	0.4186/0.0195	0.5097/-0.0042	0.3390/0.2763	0.2621/0.1742	0.3332/0.2704	0.5714/0.3062	0.5700/0.3026	0.5734/0.3227
				SRM + FFN w/ score_refine					
mDeBERTa	0.5681/0.3224	0.5364/0.3209	0.5706/0.3190	0.3277/0.2490	0.3364/0.1634	0.3052/0.2231	0.5841/0.2606	0.5735/0.2874	0.5803/0.3015
RemBert	0.5419/0.2808	0.4226/0.2376	0.5268/0.2688	0.2839/0.1695	0.2818/0.1663	0.3346/0.1743	0.5833/0.3086	0.5901/0.3194	0.5848/0.3254
InfoXLM	0.5062/0.2516	0.4300/0.2761	0.5072/0.2882	0.3256/0.2746	0.2677/0.2480	0.3059/0.2784	0.5679/0.3008	0.5753/0.3009	0.5809/0.3108

Table 8: Spearman correlation and MCC on Dev of multitask training on sentence and word levels of sentence regression module(SRM) with tagging modules(BiLSTM/FFN) and score refinement(optional)

SJTU-MTLAB's Submission to the WMT23 Word-Level Auto Completion Task

Xingyu Chen
Shanghai Jiao Tong University
galaxychen@sjtu.edu.cn

Rui Wang
Shanghai Jiao Tong University
wangrui12@sjtu.edu.cn

Abstract

Word-level auto-completion (WLAC) plays a crucial role in Computer-Assisted Translation. In this paper, we describe the SJTU-MTLAB's submission to the WMT23 WLAC task. We propose a joint method to incorporate the machine translation task to the WLAC task. The proposed approach is general and can be applied to various encoder-based architectures. Through extensive experiments, we demonstrate that our approach can greatly improve performance, while maintaining significantly small model sizes.

1 Introduction

In recent years, more and more researchers have studied computer-aided translation (CAT) that aims to assist human translators to translate the input text (Alabau et al., 2014; Knowles and Koehn, 2016; Hokamp and Liu, 2017; Santy et al., 2019; Huang et al., 2021; Weng et al., 2019). The word-level auto-completion (WLAC) task (Casacuberta et al., 2022) is the core function of CAT, which involves predicting the word being typed by the translator given the translation context, as illustrated in Figure 1. Effective auto-completion has the potential to reduce keystrokes by at least 60% during the translation process (Langlais et al., 2000). A user survey indicates that 90.2% of participants find the word-level auto-suggestion feature helpful (Moslem et al., 2022). Therefore, WLAC plays an important role in CAT.

There are many existing methods for modeling WLAC, and they mainly differ in model architectures (Li et al., 2021; Yang et al., 2022b; Moslem et al., 2022; Yang et al., 2022a; Ailem et al., 2022). For example, Li et al. (2021); Yang et al. (2022a) design a BERT-like architecture to directly predict the target word while Yang et al. (2022b) employ a model similar to the auto-regressive NMT to predict the BPE tokens of the target word.

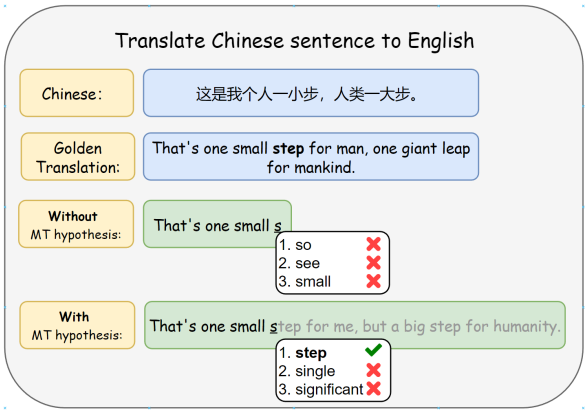


Figure 1: An example of word-level auto completion. Assume the human translator is going to input the *Golden Translation*. The auto-completion suggests the possible word candidates given the typed characters. It can be more accurate with the help of translation hypothesis from MT models.

The WLAC task comes from a real translation scenario: a human translator is translating a source sentence, who has already translated part of the sentence, and is typing a new word. The input contains three parts: the source sentence **s**, the partial translation **c**, and the typed sequence **t**. The WLAC task is to predict the word **w** that the translator is going to input (Li et al., 2021; Casacuberta et al., 2022). Rooted in the translation natural, we consider a fundamental question: what defines a correct word **w**? Theoretically, a good **w** should appear in the reference translation, as illustrated in Figure 1. Therefore, we attempt to incorporate knowledges from machine translation into the WLAC task. We presents two novel approach to enhance WLAC systems, called joint-inference and joint-training, to combine the MT task and the WLAC task during inference and training, respectively.

The effectiveness of our proposed method is validated through experiments conducted on the four language directions of the WLAC shared task

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 872–876
December 6–7, 2023.

in WMT2023 (§4). Remarkably, our approach achieves substantial improvements across two distinct backbone models.

2 Backbone Models for WLAC

In this section, we introduce two types of backbone models for the WLAC task. These backbone models serve as the foundation for our proposed techniques and experiments in subsequent sections.

Word-level Model The first backbone is called All-In-One Encoder (AIOE), which adopts a BERT-like(Devlin et al., 2019) Transformer Encoder architecture for word prediction similar to Li et al. (2021). The AIOE takes the concatenation of the source sentence, context, and typed sequence as its input. The input format is: $\boldsymbol{s}$ <*sep*> $\boldsymbol{c}_l$ <*tip*> $\boldsymbol{t}$ <*mask*> $\boldsymbol{c}_r$, where $\mathbf{c}_l$ is the left context to the input and $\mathbf{c}_r$ is the right context. Specifically, we append a <*mask*> token at the end of the typed sequence and leverage the final hidden state of the <*mask*> token for word prediction.

Despite its simplicity and efficiency, the AIOE model suffers from the out-of-vocabulary (OOV) problem, which can significantly hinder its performance. To this end, we introduce a variance of AIOE model that predicts word in sub-word level.

Sub-word-level Model Extending the word-level AIOE model to sub-word-level is straightforward: we consider the task of predicting a sequence of sub-words as a generation problem, and introduce a Transformer Decoder to the AIOE model to perform the generation. We use Byte-Pair Encoding (BPE) (Sennrich et al., 2016) for the sub-word tokenization, and call this model *AIOE-BPE*.

Due to the difficulty of labeling the WLAC data, we generate training data from parallel corpus for training the WLAC models, following the standard practice (Li et al., 2021; Casacuberta et al., 2022).

3 Enhancing WLAC by incorporating MT task

In this section, we propose two different approaches to improve the WLAC task.

3.1 Joint Inference with MT Model

This approach is to jointly consider the WLAC predictions and machine translation results during inference. We begin by generating the top-k predictions from the WLAC model. Next, we examine each word in the predictions and check if it is included in the translation. The first word in the top-k list that exists in the translation is selected as the final prediction. This strategy manually align the prediction with translation in a flexible way: the choice of WLAC model and translation model is arbitrary. The final performance is closely related to the choices of models.

However, this approach heavily relies on the quality of translation. A preliminary analysis show that for a naive MT model, only 44.6% of the WLAC labels exist in the translation. One possible solution is to enhance the input of MT model. We propose a *Context MT* model, which takes additional translation context and typed sequence as input, and generates full target sentence. The input of *Context MT* is the same as WLAC, so it's a better approximation of the golden translation model.

3.2 Joint Training with MT Task

One drawback of joint inference method is that the WLAC model isn't aware of the translation task during training, which means that the top-k predictions may deviate from the ground truth. To overcome this limitation, we propose a joint training approach, wherein the WLAC model and the MT model are trained together using a shared backbone encoder. Specifically, we extend the backbone model by introducing an MT decoder, transforming the whole model into an MT model. Here the MT model is the same as *Context MT* model described in §3.1. We define the training loss of the joint training model as the combination of the WLAC loss and the translation loss, represented as follows:

$$\mathcal{L} = \alpha \cdot L_{\text{WLAC}} + (1 - \alpha) \cdot L_{\text{MT}}, \qquad (1)$$

where α is a hyper-parameter controlling the balance between the two losses. To enhance the interaction between two tasks, we also share the final word prediction layer between the backbone model and the decoder. As described in section 4.1, the training data of WLAC is generated from parallel corpus, so there will be a full agreement between WLAC label and ground truth translation. This agreement enables the WLAC model to learn how to accurately predict words within the translations. Besides, the MT model can learn to generate translations based on the context provided by the WLAC predictions. By jointly training the two models, we enable them to mutually benefit from each other's knowledge and improve their respective tasks.

Model	#Parameters	zh-en	en-zh	en-de	de-en
AIOE	80M	46.71	54.82	51.75	50.64
AIOE-BPE	74M	50.79	53.48	57.23	61.96
AIOE+Joint Training	80M(105M)	51.40	58.70	56.22	54.57
AIOE-BPE+Joint Training	74M(100M)	**56.93**	**61.16**	**67.27**	**68.16**

Table 1: Experiment results on WMT23 WLAC test set. Results are reported as accuracy. The number of parameters in brackets means parameters in training stage.

The key advantage of joint training is that once the training is completed, we can only keep the backbone model and discard the MT decoder. Note that the backbone encoder can receive optimization signals from both the WLAC task and the translation task, so the backbone model has acquired the skill to agree with translation during training process. This enables us to maintain the agreement capabilities while preserving a small and efficient inference model.

4 Experiment

4.1 Datasets

We conduct evaluations of our model on two language pairs: English-Chinese and English-German. The zh-en dataset we used is the UN Parallel Corpus V1.0 from WMT17. For en-de, we use the training data from WMT14. We adopt the following strategy on parallel sentences to generate WLAC training data: firstly, we sample a target word **w** from the target language sentence, then we sample spans respectively from the left and right context of the target word, denoted as $\mathbf{c}_l$ and $\mathbf{c}_r$. Additionally, we sample a typed sequence from the target word. To sample typed sequence from Chinese words we use the pypinyin[1] tool. All models are trained on the generated training data, with data generated from the test set of WMT21 translation task serving as the validation set. For evaluation, we utilize the test set from the WMT22 WLAC shared task.

4.2 Experiment Details

For all AIOE model, we use a Transformer Encoder for 6 layers. The embedding size is 512, the dimension for feed-forward layer is 2048. Each layer has 8 attention heads. For AIOE-BPE model, we additionally add a Transformer Decoder with 6 layers. The MT decoder for joint training models are also 6 layers.

[1]https://github.com/mozillazg/python-pinyin

For AIOE model, we use a joint-vocabulary with the size of 120000. For AIOE-BPE model, the vocabulary size is 66630 for English-Chinese pair and 59918 for English-German pair.

The learning rate for training is 5e-4. We optimize the model for 200000 steps with a batch size of 32000 tokens. We average five checkpoints for better performance.

4.3 Comparison among Joint Methods

We firstly compare the performance of joint inference method and joint training method. For joint inference method, we use the word-level backbone AIOE model for the WLAC model, and consider two kinds of machine translation model: translation model trained on parallel corpus (MT) and translation model trained on WLAC input and translation output (*Context MT*). For the joint training method, we use AIOE-Joint model. All the experiments are conduct in zh-en direction. We conduct preliminary experiments on the WLAC22 test set and the result is reported in Table 2.

Method	Acc.
AIOE	53.87
AIOE+MT	54.20
AIOE+CMT	56.01
AIOE+JT	59.75

Table 2: Comparison of joint-methods. *Acc.* is the accuracy of WLAC22 task. AIOE+MT and AIOE+CMT is joint-inference method combined with different MT models. AIOE+JT is the joint training method.

It is observed that joint inference methods greatly outperform the baseline model, and the joint training method further improves the performance. The Context MT model is better than normal MT model for joint-inference, suggesting that more translation context is beneficial for the WLAC prediction. However, the overall performance of joint-inference is hindered by the quality of MT models, and the joint-training method can incorporate MT

Model	#Parameters	zh-en	en-zh	en-de	de-en
GWLAN(Li et al., 2021)	105M	51.11	48.90	40.69	53.87
HW-TSC(Yang et al., 2022b)	526M	59.40	-	63.82	62.06
AIOE+Joint Training	80M(105M)	59.75	56.59	44.67	62.77
AIOE-BPE+Joint Training	74M(100M)	**61.08**	**58.09**	**64.59**	**66.91**

Table 3: Experiment results on WMT22 WLAC test set. We implement the GWLAN model report the performance. The scores of HW-TSC model are copied from Yang et al. (2022b)

.

knowledge with WLAC more effectively. Based on these findings, we only focus on the joint training method for the subsequent experiments.

4.4 Main Results

The evaluation result on WLAC shared task is reported on Table 1. Our BPE-level methods have obtained better performance than word-level model except for en-zh, which indicates the word-level model may suffer from the OOV problem. No matter which backbone is used, our joint training method can greatly improve the backbone performance, indicating that our method is a general framework and has the potential to be applied to more encoder based models. Another obvious advantage of our model is its superior parameter efficiency. Our AIOE-BPE+Joint Training model achieves the best performance with only 100M training parameters and 74M parameters for inference.

4.5 Comparison with other models

We further compare our methods with existing systems. The experiment result on the WLAC22 test set is shown in Table 3. Compared to HW-TSC, our word-level methods have obtained better performance on zh-en and de-en. One exception is en-de, the word-level model performed badly because it suffers from OOV problem, where about 17% labels are OOV. After replacing the backbone with BPE-level model, our method show superior performance in all directions, while maintaining a much smaller size.

4.6 The impact of MT task

The influence of the hyper-parameter α on the model performance, as outlined in equation 1, directly reflects the impact of translation task. By setting α to 0, the model is essentially a translation model with additional context input. If $\alpha = 1$, the model corresponds to the AIOE model without joint training. In Figure 2, we present the accuracy achieved at varying values of α. Notably, as α increases from 0 to 0.75, the accuracy increases rapidly. This observation highlights the difference between the translation task and the WLAC task, emphasizing the necessity of optimizing the model specifically for the WLAC task to achieve better performance. Interestingly, even with α set to 0.99, the performance remains comparable to the best achieved performance. This finding is remarkable, as it suggests that even a small signal from the translation task can greatly enhance the WLAC task's performance when compared to the model with α set to 1. Consequently, our proposed joint training method effectively integrates the translation task into the WLAC task, resulting in substantial improvements.

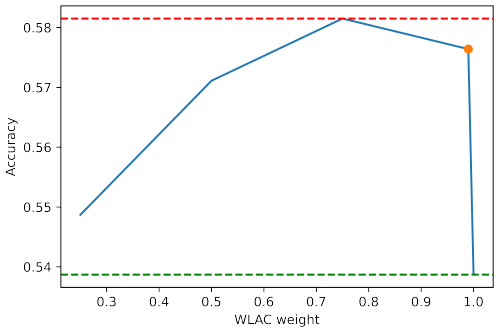


Figure 2: The impact of different α on the AIOE accuracy. Red dashed line is the best performance and the green represents the accuracy without joint training.

5 Conclusion

This paper proposes an effective approach to improve WLAC performance by combining the MT task and the WLAC task. We inject the translation knowledge into the WALC model by jointly train the two tasks. Extensive experiments show that the proposed approach surpasses several strong baselines with much smaller model size.

Acknowledgments

Xingyu and Rui are with MT-Lab, Department of Computer Science and Engineering, School of Electronic Information and Electrical Engineering, and also with the MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200204, China. Xingyu is mainly supported by Tencent Rhinobird Fund (RBFR2023012) . Rui is partially supported by the General Program of National Natural Science Foundation of China (6217020129), Shanghai Pujiang Program (21PJ1406800), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102),CCF-Baidu Open Fund (F2022018), and the Alibaba-AIR Program (22088682)

References

Melissa Ailem, Jingshu Liu, Jean-Gabriel Barthélemy, and Raheel Qader. 2022. Lingua custodia's participation at the wmt 2022 word-level auto-completion shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1170–1175.

Vicent Alabau, Christian Buck, Michael Carl, Francisco Casacuberta, Mercedes García-Martínez, Ulrich Germann, Jesús González-Rubio, Robin Hill, Philipp Koehn, Luis A Leiva, et al. 2014. Casmacat: A computer-assisted translation workbench. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*.

Francisco Casacuberta, George Foster, Guoping Huang, Philipp Koehn, Geza Kovacs, Lemao Liu, Shuming Shi, Taro Watanabe, and Chengqing Zong. 2022. Findings of the word-level autocompletion shared task in wmt 2022. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 812–820.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. *arXiv preprint arXiv:1704.07138*.

Guoping Huang, Lemao Liu, Xing Wang, Longyue Wang, Huayang Li, Zhaopeng Tu, Chengyan Huang, and Shuming Shi. 2021. Transmart: A practical interactive machine translation system. *arXiv preprint arXiv:2105.13072*.

Rebecca Knowles and Philipp Koehn. 2016. Neural interactive translation prediction. In *12th Conferences of the Association for Machine Translation in the Americas: MT Researchers' Track, AMTA 2016, Austin, TX, USA, October 28 - November 1, 2016*, pages 107–120. The Association for Machine Translation in the Americas.

Philippe Langlais, George Foster, and Guy Lapalme. 2000. Transtype: a computer-aided translation typing system. In *ANLP-NAACL 2000 Workshop: Embedded Machine Translation Systems*.

Huayang Li, Lemao Liu, Guoping Huang, and Shuming Shi. 2021. GWLAN: General word-level Autocompletion for computer-aided translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online.

Yasmin Moslem, Rejwanul Haque, and Andy Way. 2022. Translation word-level auto-completion: What can we achieve out of the box? *arXiv preprint arXiv:2210.12802*.

Sebastin Santy, Sandipan Dandapat, Monojit Choudhury, and Kalika Bali. 2019. INMT: interactive neural machine translation prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019 - System Demonstrations*, pages 103–108. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Rongxiang Weng, Hao Zhou, Shujian Huang, Lei Li, Yifan Xia, and Jiajun Chen. 2019. Correct-and-memorize: Learning to translate from interactive revisions. *arXiv preprint arXiv:1907.03468*.

Cheng Yang, Siheng Li, Chufan Shi, and Yujiu Yang. 2022a. Iigroup submissions for wmt22 word-level autocompletion task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*.

Hao Yang, Hengchao Shang, Zongyao Li, Daimeng Wei, Xianghui He, Xiaoyu Chen, Zhengzhe Yu, Jiaxin Guo, Jinlong Yang, Shaojun Li, et al. 2022b. Hw-tsc's submissions to the wmt22 word-level auto completion task. In *Proceedings of the Seventh Conference on Machine Translation Shared Tasks Papers. Association for Computational Linguistics*.

PRHLT's Submission to WLAC 2023

Ángel Navarro[1] and **Miguel Domingo**[1,2] and **Francisco Casacuberta**[1,2]
[1]PRHLT Research Center
Universitat Politècnica de València, Spain
{annamar8, midobal, fcn}@prhlt.upv.es
[2]ValgrAI - Valencian Graduate School and Research Network for Artificial Intelligence,
Camí de Vera s/n, 46022 Valencia, Spain

Abstract

This paper describes our submission to the *Word-Level AutoCompletion* shared task of WMT23. We participated in the English–German and German–English categories. We extended our last year segment-based interactive machine translation approach to address its weakness when no context is available. Additionally, we fine-tune the pre-trained mT5 large language model to be used for autocompletion.

1 Introduction

Despite its improvement in recent years with the emergence of neural machine translation (NMT), machine translation (MT) still cannot assure high-quality translations for all tasks (Toral, 2020). As a consequence, it is critical for professional translators to manually validate the translations generated by the NMT system for those scenarios with rigorous translation quality requirements. Computer-aided translation (CAT) tools emerged to improve the validation and editing process carried out by translators. With the aim of reducing the human effort of correcting the automatic translations, researchers approached CAT tools from many directions. Among CAT tools such as translation memory (Zetzche, 2007), augmented translation (Lommel, 2018) and terminology management (Verplaetse and Lambrechts, 2019); we can find autocompletion tools, which help professional translators by providing new partial translations according to the validated parts they have supplied to the system.

Word level autocompletion (WLAC) (Lin et al., 2021) was introduced as a shared task in WMT22 (Casacuberta et al., 2022). Its aim is to complete a target word given a source sentence, a sequence of characters typed by the human translator and a translation context. Four types of context are possible:

Zero-contex: no context is given.

Suffix: a sequence of translated words located after the word to autocomplete.

Prefix: a sequence of translated words located prior to the word to autocomplete.

Bi-contex: A combination of the *suffix* and the *prefix* type. That is, there is a sequence of translated words located after the word to autocomplete, and a sequence of translated words located prior to the word to autocomplete.

Note that, in all cases, the word to autocomplete is not necessarily consecutive to these contexts.

Approaches to WLAC include modeling the task as a structured prediction (generation) task (Yang et al., 2022b; Ailem et al., 2022), modeling it as a segment-based interactive machine translation (IMT) task (Navarro et al., 2022), using pre-trained NMT models and available libraries (Moslem et al., 2022), and using a generator-reranker framework (Yang et al., 2022a).

In this work, we extended the segment-based IMT approach from Navarro et al. (2022) by adding a module based on a statistical dictionary that tackles zero-context completions—which are the cases in which, not having any feedback, the IMT system performs at its worst. Additionally, since this year edition allowed the use of pre-trained large language model (LLM), we experimented using the mT5 model (Xue et al., 2021).

2 Segment-based interactive machine translation

Segment-based IMT establishes a framework in which a human translator works together with the MT system to produce the final translation. This collaboration starts with the system proposing an initial translation hypothesis y_1^I of length I. Then, the user reviews this hypothesis and validates those sequence of words which they consider

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 877–881
December 6–7, 2023.

to be correct ($\tilde{\mathbf{f}}_1, \ldots, \tilde{\mathbf{f}}_N$; where N is the number of non-overlapping validated segments). After that, they are able to merge two consecutive segments $\tilde{\mathbf{f}}_i$, $\tilde{\mathbf{f}}_{i+1}$ into a new one. Finally, they correct a word—which introduces a new one-word validated segment, $\tilde{\mathbf{f}}_i$, which is inserted in $\tilde{\mathbf{f}}_1^N$. This correction can also consist in a partially typed word $\tilde{\mathbf{f}}'_i$, in which case the system would complete it as part of its prediction.

The system's reacts to this user feedback by generating a sequence of new translation segments $\widehat{\mathbf{g}}_1^N = \widehat{\mathbf{g}}_1, \ldots, \widehat{\mathbf{g}}_N$; where each $\widehat{\mathbf{g}}_n$ is a subsequence of words in the target language. This sequence complements the user's feedback to conform the new hypothesis:

$$\begin{cases} \hat{y}_1^I = \tilde{\mathbf{f}}_1, \widehat{\mathbf{g}}_1, \ldots, \tilde{\mathbf{f}}'_i \widehat{\mathbf{g}}_i, \ldots, \tilde{\mathbf{f}}_N, \widehat{\mathbf{g}}_N \text{ if } \tilde{\mathbf{f}}'_i \in \tilde{\mathbf{f}}_1^N \\ \hat{y}_1^I = \tilde{\mathbf{f}}_1, \widehat{\mathbf{g}}_1, \ldots, \tilde{\mathbf{f}}_N, \widehat{\mathbf{g}}_N \text{ otherwise} \end{cases} \tag{1}$$

The word probability expression for the words belonging to a validated segment $\tilde{\mathbf{f}}_n$ was formalized by Peris et al. (2017) as:

$$p(y_{i_n+i'} \mid y_1^{i_n+i'-1}, x_1^J, f_1^N; \Theta) = \mathbf{y}_{i_n+i'}^\top \mathbf{p}_{i_n+i'}, \quad 1 \leq i' \leq \hat{l}_n \tag{2}$$

where l_n is the size of the non-validated segment generated by the system, which is computed as follows:

$$\hat{l}_n = \arg\max_{0 \leq l_n \leq L} \frac{1}{l_N + 1} \sum_{i'=i_n+1}^{i_n+l_n+1} \log p(y_{i'} \mid y_1^{i'-1}, x_1^J; \Theta) \tag{3}$$

3 Approaches

In this work, we extended Navarro et al. (2022)'s segment-based IMT approach by adding a new module that handles zero-context completions, which are harder for the IMT system to deal with (since there is no user feedback).

Additionally, we designed a new approach based on the mT5 LLM (Xue et al., 2021).

3.1 Segment-based IMT

Given a source sentence x_1^J, a sequence of typed characters $s_1^K = s_1, \ldots, s_K$ and a context $\mathbf{c} = \{\mathbf{c}_\text{l}, \mathbf{c}_\text{r}\}$, where $\mathbf{c}_\text{l} = c_{\text{l}1}, \ldots, c_{\text{l}S}$ and $\mathbf{c}_\text{r} = c_{\text{r}1}, \ldots, c_{\text{r}R}$; WLAC aims to autocomplete s_1^K to conform the word $w_1^W = s_1, \ldots, s_K, w_{K+1}, \ldots, w_W$. If we consider the context as the sequence of segments validated by the user ($\tilde{\mathbf{f}}_1^N = \mathbf{c}_\text{l}, \mathbf{c}_\text{r}$) and the sequence s_1^K as the partially-typed word correction (which would be inserted in $\tilde{\mathbf{f}}_1^N$ as a new one-word validated segment; leading to $\tilde{\mathbf{f}}_1^N = \mathbf{c}_\text{l}, s_1^K, \mathbf{c}_\text{r}$), we can view WLAC as a simplification of segment-based IMT. With that in mind, we can rewrite Eq. (1) as:

$$\hat{y}_1^I = \mathbf{c}_\text{l}, \widehat{\mathbf{g}}_1, s_1^K \widehat{\mathbf{g}}_2, \mathbf{c}_\text{r}, \widehat{\mathbf{g}}_3 \tag{4}$$

which, knowing that the prediction of the partially-typed correction corresponds to the first word of $\widehat{\mathbf{g}}_2$, can be rewritten as:

$$\hat{y}_1^I = \mathbf{c}_\text{l}, \widehat{\mathbf{g}}_1, s_1^K w_{K+1}^W, \widehat{\mathbf{g}}'_2, \mathbf{c}_\text{r}, \widehat{\mathbf{g}}_3 \tag{5}$$

Therefore, we can obtain the autocompleted word ($w_1^W = s_1^K w_{K+1}^W$) by performing a single step of the segment-based IMT protocol, discarding the rest of the translation prediction.

Zero-context

Since the core idea of IMT is reacting to a user feedback, not having any context results in the segment-based IMT approach performing at its worst. Thus, in this work we decided to create a special module dedicated to perform this kind of completion, using a variation of a statistical dictionary.

To that end, we computed *IBM's model 1* (Och and Ney, 2003) to obtain word alignments from source and target of the training set. Then, for each source word x_j, we compute the most probable translation t_a that starts with the sequence to complete s_1^K ($t_a = s_1, \ldots, s_K, t_{K+1}, \ldots, t_T$):

$$\hat{t}_j = \arg\max_{t_a} p(t_a | x_j) \tag{6}$$

where t_a belongs to the set of target words aligned with x_j that starts with s_1^K; and $p(t_a|x_j)$ is the alignment probability given by *IBM's model 1*.

Finally, we obtain the autocompleted word w_1^W as the most probable translation:

$$w_1^W = \arg\max_{t_1^J} p(t_1^J | x_1^J) \tag{7}$$

3.2 mT5

mT5 (Xue et al., 2021) is a multilingual variant of T5 (Raffel et al., 2020), pre-trained on a new Common Crawl-based dataset covering 101 languages. We choose to use this LLM since it has been pre-trained without any supervised training and, thus,

```
{
"src": "Indonesischer Lehrerin droht Haftstrafe wegen Dokumentation von sexueller Belästigung",
"context_type": "prefix",
"target": "school",
"typed_seq": "sch",
"left_context": "Indonesian",
"right_context": "",
"segment_id": "ref0"
},
```


(a) Original sentence in json format.

Indonesischer Lehrerin droht Haftstrafe wegen Dokumentation von sexueller Belästigung ||| Indonesian ||| ||| sch

(b) mT5 source sentence.

school

(c) mT5 target sentence.

Figure 1: Example of adapting a training sentence for fine-tuning mT5.

can be easily adapted to any downstream task by simply fine-tuning the model.

Therefore, this approach consists in fine-tuning mT5 for WLAC. To do so, we created a new parallel dataset in which source sentences are the concatenation of the original source sentence, the left context, the right context and the typed sequence (using a special token as a delimiter); and target sentences are the autocompletion. Fig. 1 shows an example.

4 Experimental setup

In this section, we present the details of our experimental session.

4.1 Evaluation

The WLAC 23 shared task selected accuracy as the automatic metric with which to report the evaluation of the different systems. This metric is computed as the total number of correctly predicted words normalized by the total number of words to complete:

$$\text{Acc} = N_{\text{match}}/N_{\text{all}} \tag{8}$$

where N_{match} is the number of predicted words that are identical to the human desired word, and N_{all} is the total number of testing words.

4.2 Corpora

We conducted our experiments using the English–German corpus provided by the organizers, which is a version of the WMT14's dataset, preprocessed by Stanford NLP Group.

Table 1: Statistics of the WLAC 2023 corpus. *Run.* stands for running, K for thousands and M for millions.

Partition	Characteristic	De	En
Training	Sentences	$4M$	
	Run. Words	$110M$	$116M$
	Vocabulary	$1.6M$	$800K$
Validation	Sentences	2000	
	Run. Words	$53K$	$53K$
	Vocabulary	$10.5K$	$7.5K$

For fine-tuning mT5 (see Section 3.2), we processed the training data using the provided script[1] in order to create the simulated data. We repeated this process multiple times to increase the number of samples. Table 2 presents the data statistics.

Table 2: Statistics of the synthetic corpus generated for fine-tuning the mT5 model. *Run.* stands for running, K for thousands and M for millions.

Partition	Characteristic	De	En
Training	Sentences	$50M$	
	Run. Words	$1627.6M$	$1677.6M$
	Vocabulary	$1.6M$	$800K$
Validation	Sentences	2000	
	Run. Words	$53K$	$53K$
	Vocabulary	$93.4K$	$144.9K$

[1] https://github.com/lemaoliu/WLAC/raw/main/scripts/generate_samples.py.

Table 3: Experimental results, measured in terms of accuracy.

Approach	Language	Overall	Prefix	Suffix	Bi-context	Zero-context
Segment-base IMT	De–En	0.400	0.453	0.151	0.395	0.570
	En-De	0.371	0.433	0.144	0.377	0.491
mT5	De–En	0.436	0.432	0.458	0.490	0.363
	En–De	0.374	0.373	0.389	0.431	0.301

4.3 Systems

The MT systems from our segment-based IMT approach were trained using *OpenNMT-py* (Klein et al., 2017). We selected a Transformer (Vaswani et al., 2017) architecture, with a word embedding size of 512. The hidden and output layers were set to 2048 and 512, respectively. Each multi-head attention layer has eight heads, and we stacked six encoder and decoder layers. We used Adam as the learning algorithm, with a learning rate of 2.0, b_1 of 0.9 and b_2 of 0.998. We set the batch size to 4096 tokens.

Additionally, we made use of the byte pair encoding (BPE) (Sennrich et al., 2016) algorithm, which was jointly trained on both languages of the dataset, applying a maximum number of 10.000 merges. Finally, we used our own implementation (based on *OpenNMT-py*) of segment-based IMT, which we adapted for WLAC. This implementation is openly available[2] for the benefit of the community.

For the mT5 approach, we made use of *HuggingFace's Transformer* (Wolf et al., 2019). Due to computing constrains, we selected *Google's mT5-base* model[3].

5 Results

Table 3 presents the official results of our approaches. We can see how both approaches yielded similar results. The main advantage of the segment-base IMT approach is that we can leverage an MT model for autocompletion by simply performing minor changes at the decoding step. However, looking at the results, while our zero-context proposal has successfully solved the problem of having no feedback, the system's performance significantly drops when the only available context is a suffix. In future works we shall address this behavior.

Regarding the mT5 approach, its main advantage is that we can adapt an already pre-trained mT5 model by simply performing fine-tuning with a WLAC dataset. With the exception of having no context, its behavior is constant for all kind of context. Additionally, it is worth remembering that we used *Google's mT5-base* model due to computing constrains. In a future work, we shall test how "bigger" mT5 models behave for this task.

6 Conclusions

In this work, we have presented our submission to WLAC shared task from WMT23. Our first proposal extended Navarro et al. (2022)'s segment-based IMT approach by adding a zero-context—based on a statistical dictionary—that handles separately the cases in which no context is given. This approach yielded satisfactory results for all cases except when the given context consists in a suffix.

Our second proposal consisted in leverage the pre-trained LLM model mT5 by performing a simple fine-tuning that enables the model to be used for WLAC, achieving satisfactory results for all type of contexts.

As a future work, we would like to study the behavior of the segment-based IMT approach when dealing with suffixes. Additionally we would like to consider the use of other LLM, as well as different versions of the mT5 model.

Acknowledgements

This work received funding from *Generalitat Valencia* under the program *CIACIF/2021/292* and from *ValgrAI* (*Valencian Graduate School and Research Network for Artificial Intelligence*). It has also been partially supported by grant *PID2021-124719OB-I00* funded by *MCIN/AEI/10.13039/501100011033* and by *European Regional Development Fund* (*ERDF*).

References

Melissa Ailem, Jingshu Liu, Jean-Gabriel Barthélemy, and Raheel Qader. 2022. Lingua custodia's

[2] https://github.com/PRHLT/OpenNMT-py/tree/word-level_autocompletion.

[3] https://huggingface.co/google/mt5-base.

participation at the WMT 2022 word-level autocompletion shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1170–1175.

Francisco Casacuberta, George Foster, Guoping Huang, Philipp Koehn, Geza Kovacs, Lemao Liu, Shuming Shi, Taro Watanabe, and Chengqing Zong. 2022. Findings of the word-level autocompletion shared task in wmt 2022. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 812–820.

G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of the Association for Computational Linguistics: System Demonstration*, pages 67–72.

Huayang Lin, Lemao Liu, Guoping Huang, and Shuming Shi. 2021. GWLAN: General word-level autocompletion for computer-aided translation. In *Proceedings of the Joint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*. In Press.

Arle Lommel. 2018. Augmented translation: A new approach to combining human and machine capabilities. In *Proceedings of the Conference of the Association for Machine Translation in the Americas. Volume 2: User Track*, pages 5–12.

Yasmin Moslem, Rejwanul Haque, and Andy Way. 2022. Translation word-level auto-completion: What can we achieve out of the box? In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1176–1181.

Ángel Navarro, Miguel Domingo, and Francisco Casacuberta. 2022. PRHLT's submission to WLAC 2022. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1182–1186.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Álvaro Peris, Miguel Domingo, and Francisco Casacuberta. 2017. Interactive neural machine translation. *Computer Speech & Language*, 45:201–220.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.

Antonio Toral. 2020. Reassessing claims of human parity and super-human performance in machine translation at wmt 2019. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 185–194.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Heidi Verplaetse and An Lambrechts. 2019. Surveying the use of CAT tools, terminology management systems and corpora among professional translators: general state of the art and adoption of corpus support by translator profile. *Parallèles*, 31(2):3–31.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Cheng Yang, Siheng Li, Chufan Shi, and Yujiu Yang. 2022a. Iigroup submissions for wmt22 word-level autocompletion task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1187–1191.

Hao Yang, Hengchao Shang, Zongyao Li, Daimeng Wei, Xianghui He, Xiaoyu Chen, Zhengzhe Yu, Jiaxin Guo, Jinlong Yang, Shaojun Li, et al. 2022b. Hw-tsc's submissions to the wmt22 word-level auto completion task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1192–1197.

Jost Zetzche. 2007. Translation memory: state of the technology. *Multilingual*, 18:34–38.

KnowComp Submission for WMT23 Word-Level AutoCompletion Task

Yi Wu[1,2], Haochen Shi[2], Weiqi Wang[2], Yangqiu Song[2]
[1]University of Wisconsin-Madison, WI, USA
[2]Department of Computer Science and Engineering, HKUST, Hong Kong SAR, China
ywu676@wisc.edu

Abstract

The NLP community has recently witnessed the success of Large Language Models (LLMs) across various Natural Language Processing (NLP) tasks. However, the potential of LLMs for word-level auto-completion in a multilingual context has not been thoroughly explored yet. To address this gap and benchmark the performance of LLMs, we propose an LLM-based system for the WMT23 Word-Level Auto-Completion (WLAC) task. Our system utilizes ChatGPT to represent LLMs and evaluates its performance in three translation directions: Chinese-English, German-English, and English-German. We also study the task under zero-shot and few-shot settings to assess the potential benefits of incorporating exemplars from the training set in guiding the LLM to perform the task. The results of our experiments show that, on average, our system attains a 29.8% accuracy on the test set. Further analyses reveal that LLMs struggle with WLAC in the zero-shot setting, but performance significantly improves with the help of additional exemplars, though some common errors still appear frequently. These findings have important implications for incorporating LLMs into computer-aided translation systems, as they can potentially enhance the quality of translations. Our codes for evaluation are available at https://github.com/ethanyiwu/WLAC.

1 Introduction

Recent advancements in machine translation, especially due to the development of transformers and pre-trained language models, have been significant (Kong and Fan, 2021; Sun et al., 2023; Mohammadshahi et al., 2022). These methods have yielded impressive results in traditional sentence-level translation tasks (Bahdanau et al., 2015). However, challenges still exist that hinder the further progress of Computer-Aided Translation (CAT) (Esplà-Gomis et al., 2022) systems. Among various components that constitute CAT, Word-level AutoCompletion (WLAC) stands out as a core function (Li et al., 2021; Casacuberta et al., 2022). As shown in Figure 1, WLAC aims to suggest the correct word translation in the target language based on a sequence of human-typed characters and bidirectional context.

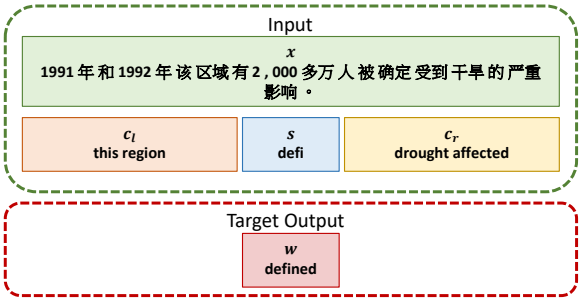


Figure 1: Illustration of the WLAC task for translating from Chinese to English, including various components involved in the translation process. The inputs consist of the source sentence x, the left context in the target sentence c_l, the right context in the target sentence c_r, and the pre-typed character sequence of the word to be predicted s. The task aims to predict the target output word w accurately.

While this task might seem straightforward for seasoned human translators, existing deep-learning approaches struggle to handle it effectively. This can be attributed to the fact that performant translation methods, which rely on pre-trained language models, cannot effectively interpret the typed sequence of characters as they are pre-trained at the token level. Other related studies have either only considered the source contextualization (Huang et al., 2015) or have been unable to handle multilingual translations effectively (Huang et al., 2018). Previous works have demonstrated that transformer-based frameworks, when trained with carefully designed masking or context transformation strategies, can efficiently tackle this task (Yang et al., 2022a,b; Navarro et al., 2022). Yet, these frameworks require extensive training, and their

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 882–889
December 6–7, 2023.

ability to transfer across different languages remains questionable.

With the recent progress made by Large Language Models (LLMs), such as ChatGPT (OpenAI, 2022), researchers have conducted extensive studies regarding their performances on various NLP tasks (Laskar et al., 2023; Liu et al., 2023; Li et al., 2023a; Yu et al., 2023). These LLMs possess advantages, such as ease of deployment and robust multilingual reasoning ability, making them particularly suitable for tasks like WLAC. However, a detailed study of LLMs' application in this context is lacking, and the systematic use of LLMs to assist CAT, especially in performing WLAC, remains unexplored.

This paper addresses these gaps by introducing a system that employs LLMs, specifically ChatGPT, for the WLAC task. ChatGPT has been chosen for its exceptional performance in general natural language tasks, negating the need for deploying or fine-tuning other models. Our approach uses a prompt-engineering-based method to convert all WLAC task inputs into comprehensible natural language sentences. Following this, we synthesize exemplars from the existing training set to facilitate in-context learning with ChatGPT (Wei et al., 2022b). The generations are subsequently parsed using carefully crafted rules to yield the final "predictions" from ChatGPT (Section 3). We then conduct comprehensive experiments using our system, covering scenarios from zero-shot to five-shots. Our experimental results indicate that our system achieves an average accuracy of 29.8% on the testing set. Through error analysis and case studies, we found that LLMs face challenges with WLAC in the zero-shot setting and identified four common types of mistakes that can be particularly addressed in the future. However, ChatGPT's performance significantly improves when provided with additional exemplars, highlighting the crucial role of in-context learning in tackling WLAC for LLMs (Section 4). We will make all codes publicly available upon acceptance of this paper.

2 Preliminaries

2.1 Task Definition

Formally, the objective of the WLAC task (Li et al., 2021) is to predict the target word w using three parts of inputs, which are denoted as the source sequence x, the human-typed characters s, and the translation context c, where $c = (c_l, c_r)$. The translation context consists of left context c_l and right context c_r, where c_l is a sub-sequence of the translated context on the left side of s, and c_r is a sub-sequence of the translated context on the right side of s. A running example is shown in Figure 1.

	Training	Validation	Test
zh-en	39,473	29,051	16,386
de-en	40,000	29,596	14,564
en-de	40,000	29,895	14,539

Table 1: Statistics regarding the number of data across three translation languages in each split.

Specifically, a notable challenge in training a masked language model for the WLAC task is the incomplete nature of the left and right contexts. These contexts may not necessarily constitute complete sentences; they can consist of partial words or even be empty. As a result, the context c and the typed sequence s do not necessarily provide a fully translated result of the source sequence. Moreover, the training data for this task does not include the complete translated result as a reference. This lack of complete supervision further complicates the establishment of robust training signals for masked language modeling (Li et al., 2023b), especially when compared to traditional translation tasks (Navarro et al., 2022).

2.2 Large Language Models

The emergence of large language models (LLMs) has recently gained the spotlight in the NLP community. GPT3.5 (Brown et al., 2020; Ouyang et al., 2022), ChatGPT (OpenAI, 2022), and LLaMA (Touvron et al., 2023) are some of the notable LLMs that have been well-developed, each boasting an exceptionally vast number of parameters. These LLMs are trained on massive corpora using advanced techniques, such as instruction tuning (Wei et al., 2022a) and reinforcement learning from human feedback (Christiano et al., 2017), on large computational infrastructures. As a result, recent studies have shown that LLMs excel at various downstream tasks, including causal reasoning and grounding (Chan et al., 2023; Wang et al., 2023c; Ou et al., 2023), commonsense reasoning (Fang et al., 2023, 2021b,a; Bian et al., 2023; Wang et al., 2023b), question-answering (Wang et al., 2023a; Qin et al., 2023), translation (Peng et al., 2023; Lu et al., 2023), and data mining tasks (Jin et al., 2023a,b,c). Since the WLAC task demands substantial reasoning and generation capabilities that

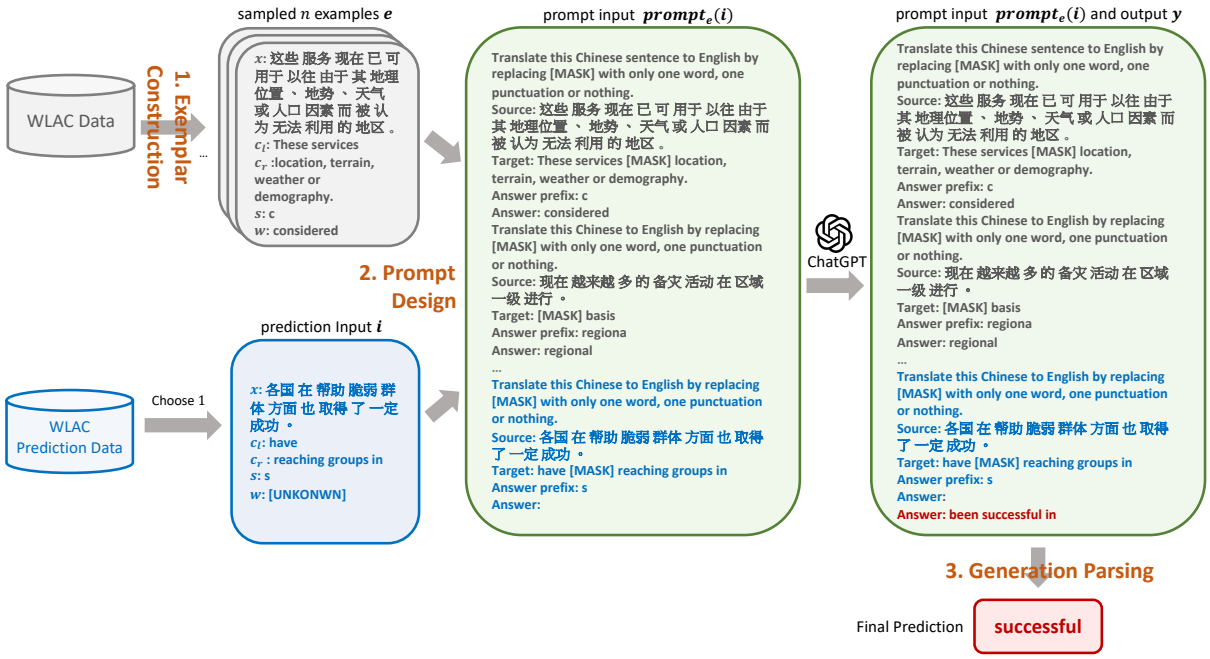


Figure 2: An overview of our framework when dealing with WLAC in translating from Chinese to English (zh-en). The input sentences (marked in blue) are concatenated with pre-constructed exemplars (marked in gray) to form a unified prompt. A large language model (ChatGPT) is then deployed to generate the response (marked in red), which is subsequently parsed to obtain the final prediction y.

rely on bidirectional contexts to accurately predict the target word, LLMs make for an ideal choice to perform WLAC due to their exceptional natural language understanding abilities and ease of deployment.

2.3 Dataset

We use the dataset provided by (Casacuberta et al., 2022) as our primary evaluation benchmark. We select three translation language pairs from the dataset: Chinese to English (zh-en), German to English (de-en), and English to German (en-de). To maintain consistency, we follow the *trn/dev/test* split released in the original dataset. Detailed statistics on the number of data are shown in Table 1.

3 Method

Figure 2 shows an overview of our framework, which consists of three steps: exemplar sampling, prompt design, and generation parsing.

3.1 Exemplar Sampling and Construction

To generate the input prompt for the LLM, we begin by employing random sampling to choose k data instances from the training split of the dataset. These selected instances serve as in-context learning exemplars. In our experiments, we explore different values of $k \in \{0, 1, 5\}$ to evaluate the performance of the LLM in both zero-shot and few-shot scenarios. The aim is to improve the model's familiarity with the task and its capacity to deliver precise answers by incorporating the provided exemplars into its learning process.

3.2 Prompt Design

We then design a natural language prompt to systematically combine both sampled exemplars and every testing data entry from the testing split, which serves as the input for the LLM. To assist the LLM in distinguishing different components of the input and the desired output, we introduce instructive tokens such as "Source" (the source sentence to be translated), "Target" (the target sentence with context c_l and c_r provided), "Answer prefix" (the pre-typed sequence s indicating the target word to be predicted at `[MASK]`), and "Answer" (representing the target prediction word w). For each testing data entry, we construct such a prompted sentence with the "Answer" for the testing entry left blank, awaiting completion. Combining all these prompted sentences creates a comprehensive paragraph of sentence input, as illustrated in Figure 2. This transforms the task into a blank-filling exercise, wherein the model fills in the missing word, the last word in our case.

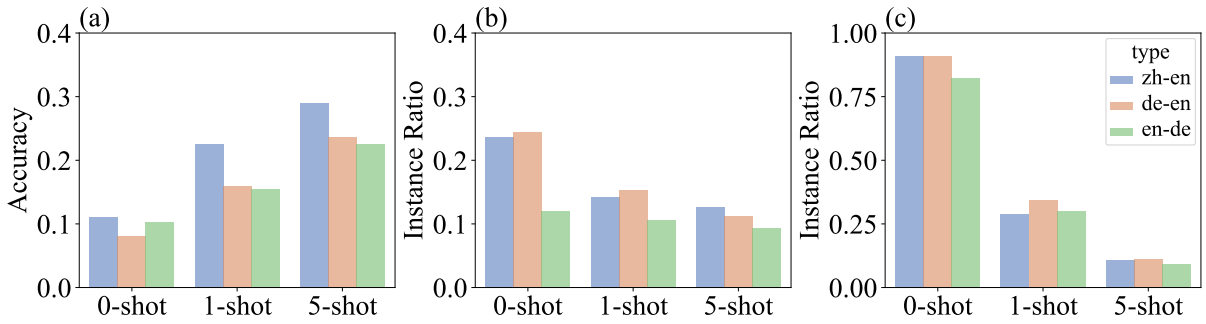


Figure 3: The fault rate and accuracy under different settings. (a) refers to the accuracy. (b) refers to the proportion of instances that don't start with the typed sequence s. (c) refers to the ratio of the generations which contain a sequence of words.

3.3 Generation Parsing

Since the generated content is in free-text form, it requires parsing to identify the predicted word as the final output of the WLAC task. The first step is to remove the cue word, such as "Answer," and separate the remaining sequence of tokens into individual words by splitting at blank spaces. To ensure that the selected word satisfies the constraint of the pre-typed sequence, we search for the first word that starts with the pre-typed sequence s as the final prediction. If such a word does not exist, the first word in the sequence is chosen as the final prediction. For example, if the generated content is "Answer: Hello World" and the pre-typed sequence is "Wo," the word "World" is the final prediction after parsing.

4 Experiments

4.1 Setup

We utilize the official ChatGPT API[1] to access the large language model for sentence completion. The model code employed is `gpt-3.5-turbo`, and the access date is July 2023. The temperature is set to 0.7 when generating to ensure consistent generation, while other control parameters are set to their default values.

4.2 Results

Our experimental results are presented in Table 2. We observe that five-shot prompting yielded the best performance, while not incorporating any training exemplar led to the worst performance. This outcome is reasonable, considering the model may not clearly understand the task objective and reasoning process. Importantly, incorporating just one additional exemplar had a significantly positive impact. This suggests that ChatGPT can quickly learn from provided exemplars and develop a sufficient understanding of the task. The accuracy improvement trends across the three translation directions are consistent, leading us to conclude that ChatGPT can achieve acceptable performance on the WLAC task with the help of training exemplars and in-context learning. However, even with five-shot prompting, the performance is only around 25% in terms of accuracy, leaving a large space for future improvements. Therefore, leveraging more advanced or meticulously designed prompts should be considered further to enhance ChatGPT's performance on the WLAC task.

Task	Shot #		
	0-shot	1-shot	5-shot
Zh - En	11.15	22.66	29.21
En - De	10.26	15.45	22.51
De - En	8.12	15.99	23.66

Table 2: Evaluation results (Accuracy %) on the testing sets of three translation directions.

4.3 Error Analysis

Upon further analysis of the results, we identify two common types of mistakes where the generated output deviates from the targeted answer. The first type of mistake occurs when the generated output fails to begin with the specified sequence s, while the second type of mistake involves the presence of multiple words in the generated output after removing the cue word. As depicted in Figure 3, the overall performance improves significantly as the number of shots increases. Nevertheless, there is a remarkably high rate of faults in generating

[1] https://chat.openai.com/

Source	Context	Generation	Target
CORRECT EXAMPLES			
据报道，受重伤的维和人员得以继续飞行，并与其他机组人员一起成功降落在北基伍省会戈马机场。	It was reported that [MASK]	Answer: seriously	seriously
他要求刚果（金）当局对这起令人发指的袭击事件展开调查，尽快将肇事者绳之以法。	The congo [MASK]	Answer: authorities	authorities
GRAMMATICAL MISTAKE			
她说："我谴责这次袭击，必须以最坚定的态度起诉犯罪者。"	attack and the perpetrators must be prosecuted with the utmost [MASK]	Answer: firmly	firmness
报告说，迫使巴勒斯坦人背井离乡的"胁迫性环境"使巴勒斯坦社会四分五裂，阻碍了自决权的实现。	" coercive environment " that forced the [MASK]	Answer: Palestinian	Palestinians
BOTH CORRECT?			
委员会呼吁联合国大会要求国际法院就占领的法律后果发表紧急咨询意见。	The [MASK]	Answer: committee	called
专家们注意到，欧盟反欺诈办公室就对独立人权组织哈克进行了审查，其结论是："没有发现受怀疑的违规和(或)欺诈行为来影响欧盟的资金"。	experts [MASK]	Answer: noticed	noted
DON'T START WITH TYPED SEQUENCE			
另据报道，在8月18日及21日，以色列国内安全局审问了7个团体中的巴勒斯坦妇女委员会联盟、独立人权组织哈克和保卫儿童 - 巴勒斯坦组织的负责人，据称还对这三人加以威胁。	It was also reported that , on 18 and 21 [MASK]	Answer: August	Actions
谭德塞说："自那时以来，世卫组织已报告了3200多例猴痘确诊病例和一例死亡，这些病例来自包括尼日利亚在内的48个国家和5个世卫组织地区。"	, WHO has reported more than 3200 confirmed [MASK]	Answer: cases	regions
GENERATE A SEQUENCE OF WORDS			
委员会成员克里斯·西多蒂（Chris Sidoti）表示，以色列政府的行动构成了一种非法占领和吞并制度，必须加以解决。	Chris Sidoti said the [MASK]	Answer: Israeli government	Israeli
委员会的成员不是联合国工作人员，他们的工作没有报酬。	The members of the Committee [MASK]	Answer: are not UN staff, their work is voluntary.	not

Table 3: Case studies of generations from ChatGPT. We select generation results from the 5-shot scenario.

word sequences that violate the instruction of using only a single word in the zero-shot approach. The transition from zero-shot to one-shot learning results in a considerable reduction in both fault rates. This indicates that the language model adheres to instructions more accurately by adding a single example. Moreover, the fault rate also further decreases in the five-shot setting.

4.4 Case Studies

To further demonstrate the difficulty of the task and the performance of ChatGPT, we select some generation results in the Chinese-English translation split, as shown in Table 3. Among these cases, we observe four types of tricky but common mistakes. Firstly, although the generated output and the target share the same semantic meaning, the generation is syntactically incorrect. We exemplify two generations that contain grammatical mistakes. Secondly, our approach may generate semantically accurate output that deviates from the target, particularly in cases where the input lacks detailed context. Moreover, the model may generate the content immediately after the context instead of following instructions to find a semantically correct word that matches the typed sequence. Finally, the model does not always comply with the instructions that require it to generate only one word. It may give a phrase or part of a sentence to combine the context into a complete sentence.

4.5 Discussions

While our system achieves acceptable performance, it falls significantly short of the performance achieved by systems last year (Casacuberta et al., 2022). This suggests that additional efforts are required to enhance ChatGPT's performance on the WLAC task, which might include: (a) Incorporating more training exemplars. For instance, increasing the number of training shots to ten or even more could be beneficial. (b) Reframing the exemplar selection problem as a subset selection problem. This approach involves selecting training exemplars based on their similarity to the testing entry or their diversity in relation to other exemplars, as proposed by Ye et al. (2023). (c) Improving the prompt to better leverage both left and right contexts. Additionally, advanced prompting techniques like chain-of-thought (Wei et al., 2022b) could be explored. (d) Incorporating external knowledge for reasoning, such as complex knowledge (Bai et al., 2023), conceptualization (He et al., 2022), and graph reasoning (Liu and Song, 2022; Liu et al., 2022, 2020).

5 Conclusions

In conclusion, this paper presents a novel LLM-prompting system to address the WLAC task. Our findings demonstrate that LLMs are highly capable

problem solvers and adept at learning in context for this particular task, albeit with performance that falls short of previous supervised learning systems. A detailed analysis uncovers several error types that contribute to the limited performance of ChatGPT. Therefore, we urge researchers to devote additional attention to the WLAC task using LLMs.

Acknowledgements

The authors would like to thank the committee of WMT2023, the organizers of the WLAC task, and the anonymous reviewers. The authors of this paper were supported by the NSFC Fund (U20B2053) from the NSFC of China, the RIF (R6020-19 and R6021-20), and the GRF (16211520 and 16205322) from RGC of Hong Kong. We also thank the support from the UGC Research Matching Grants (RMGS20EG01-D, RMGS20CR11, RMGS20CR12, RMGS20EG19, RMGS20EG21, RMGS23CR05, RMGS23EG08).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jiaxin Bai, Xin Liu, Weiqi Wang, Chen Luo, and Yangqiu Song. 2023. Complex query answering on eventuality knowledge graph with implicit logical constraints. *CoRR*, abs/2305.19068.

Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, and Ben He. 2023. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. *CoRR*, abs/2303.16421.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Francisco Casacuberta, George Foster, Guoping Huang, Philipp Koehn, Geza Kovacs, Lemao Liu, Shuming Shi, Taro Watanabe, and Chengqing Zong. 2022. Findings of the word-level autocompletion shared task in WMT 2022. In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 812–820. Association for Computational Linguistics.

Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2023. Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations. *CoRR*, abs/2304.14827.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307.

Miquel Esplà-Gomis, Víctor M. Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. 2022. Cross-lingual neural fuzzy matching for exploiting target-language monolingual corpora in computer-aided translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 7532–7543. Association for Computational Linguistics.

Tianqing Fang, Quyet V. Do, Sehyun Choi, Weiqi Wang, and Yangqiu Song. 2023. CKBP v2: An expert-annotated evaluation set for commonsense knowledge base population. *CoRR*, abs/2304.10392.

Tianqing Fang, Weiqi Wang, Sehyun Choi, Shibo Hao, Hongming Zhang, Yangqiu Song, and Bin He. 2021a. Benchmarking commonsense knowledge base population with an effective evaluation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8949–8964. Association for Computational Linguistics.

Tianqing Fang, Hongming Zhang, Weiqi Wang, Yangqiu Song, and Bin He. 2021b. DISCOS: bridging the gap between discourse knowledge and commonsense knowledge. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 2648–2659. ACM / IW3C2.

Mutian He, Tianqing Fang, Weiqi Wang, and Yangqiu Song. 2022. Acquiring and modelling abstract commonsense knowledge via conceptualization. *CoRR*, abs/2206.01532.

Guoping Huang, Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2015. A new input method for human translators: Integrating machine translation effectively and imperceptibly. In *Proceedings of the Twenty-Fourth*

International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015, pages 1163–1169. AAAI Press.

Yafang Huang, Zuchao Li, Zhuosheng Zhang, and Hai Zhao. 2018. Moon IME: neural-based chinese pinyin aided input method with customizable association. In *Proceedings of ACL 2018, Melbourne, Australia, July 15-20, 2018, System Demonstrations*, pages 140–145. Association for Computational Linguistics.

Yiqiao Jin, Yunsheng Bai, Yanqiao Zhu, Yizhou Sun, and Wei Wang. 2023a. Code recommendation for open source software developers. In *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 1324–1333. ACM.

Yiqiao Jin, Yeon-Chang Lee, Kartik Sharma, Meng Ye, Karan Sikka, Ajay Divakaran, and Srijan Kumar. 2023b. Predicting information pathways across online communities. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pages 1044–1056. ACM.

Yiqiao Jin, Xiting Wang, Yaru Hao, Yizhou Sun, and Xing Xie. 2023c. Prototypical fine-tuning: Towards robust performance under varying data sizes. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 12968–12976. AAAI Press.

Yawei Kong and Kai Fan. 2021. Probing multi-modal machine translation with pre-trained language model. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 3689–3699. Association for Computational Linguistics.

Md. Tahmid Rahman Laskar, M. Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy X. Huang. 2023. A systematic study and comprehensive evaluation of chatgpt on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 431–469. Association for Computational Linguistics.

Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, and Yangqiu Song. 2023a. Multi-step jailbreaking privacy attacks on chatgpt. *CoRR*, abs/2304.05197.

Haoran Li, Mingshi Xu, and Yangqiu Song. 2023b. Sentence embedding leaks more information than you expect: Generative embedding inversion attack to recover the whole sentence. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14022–14040. Association for Computational Linguistics.

Huayang Li, Lemao Liu, Guoping Huang, and Shuming Shi. 2021. GWLAN: general word-level autocompletion for computer-aided translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4792–4802. Association for Computational Linguistics.

Xin Liu, Jiayang Cheng, Yangqiu Song, and Xin Jiang. 2022. Boosting graph structure learning with dummy nodes. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 13704–13716. PMLR.

Xin Liu, Haojie Pan, Mutian He, Yangqiu Song, Xin Jiang, and Lifeng Shang. 2020. Neural subgraph isomorphism counting. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 1959–1969. ACM.

Xin Liu and Yangqiu Song. 2022. Graph convolutional networks with dual message passing for subgraph isomorphism counting and matching. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 7594–7602. AAAI Press.

Xin Liu, Yuan Tan, Zhenghang Xiao, Jianwei Zhuge, and Rui Zhou. 2023. Not the end of story: An evaluation of chatgpt-driven vulnerability description mappings. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3724–3731. Association for Computational Linguistics.

Qingyu Lu, Baopu Qiu, Liang Ding, Liping Xie, and Dacheng Tao. 2023. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt. *CoRR*, abs/2303.13809.

Alireza Mohammadshahi, Vassilina Nikoulina, Alexandre Berard, Caroline Brun, James Henderson, and Laurent Besacier. 2022. Small-100: Introducing shallow multilingual machine translation model for low-resource languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 8348–8359. Association for Computational Linguistics.

Ángel Navarro, Miguel Domingo, and Francisco Casacuberta. 2022. Prhlt's submission to WLAC 2022. In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages

1182–1186. Association for Computational Linguistics.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. *OpenAI*.

Jiefu Ou, Adithya Pratapa, Rishubh Gupta, and Teruko Mitamura. 2023. Hierarchical event grounding. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 13437–13445. AAAI Press.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of chatgpt for machine translation. *CoRR*, abs/2303.13780.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *CoRR*, abs/2302.06476.

Simeng Sun, Maha Elbayad, Anna Sun, and James Cross. 2023. Efficiently upgrading multilingual machine translation models to support more languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 1505–1519. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Weiqi Wang, Tianqing Fang, Wenxuan Ding, Baixuan Xu, Xin Liu, Yangqiu Song, and Antoine Bosselut. 2023a. CAR: conceptualization-augmented reasoner for zero-shot commonsense question answering. *CoRR*, abs/2305.14869.

Weiqi Wang, Tianqing Fang, Baixuan Xu, Chun Yi Louis Bo, Yangqiu Song, and Lei Chen. 2023b. CAT: A contextualized conceptualization and instantiation framework for commonsense reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13111–13140. Association for Computational Linguistics.

Zhaowei Wang, Quyet V. Do, Hongming Zhang, Jiayao Zhang, Weiqi Wang, Tianqing Fang, Yangqiu Song, Ginny Y. Wong, and Simon See. 2023c. COLA: contextualized commonsense causal reasoning from the causal inference perspective. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5253–5271. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Cheng Yang, Siheng Li, Chufan Shi, and Yujiu Yang. 2022a. IIGROUP submissions for WMT22 word-level autocompletion task. In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 1187–1191. Association for Computational Linguistics.

Hao Yang, Hengchao Shang, Zongyao Li, Daimeng Wei, Xianghui He, Xiaoyu Chen, Zhengzhe Yu, Jiaxin Guo, Jinlong Yang, Shaojun Li, Yuanchang Luo, Yuhao Xie, Lizhi Lei, and Ying Qin. 2022b. Hw-tsc's submissions to the WMT22 word-level auto completion task. In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 1192–1197. Association for Computational Linguistics.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2023. Compositional exemplars for in-context learning. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 39818–39833. PMLR.

Changlong Yu, Weiqi Wang, Xin Liu, Jiaxin Bai, Yangqiu Song, Zheng Li, Yifan Gao, Tianyu Cao, and Bing Yin. 2023. Folkscope: Intention knowledge graph construction for e-commerce commonsense discovery. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1173–1191. Association for Computational Linguistics.

Terminology-Aware Translation with Constrained Decoding and Large Language Model Prompting

Nikolay Bogoychev* **Pinzhen Chen***
School of Informatics, University of Edinburgh
n.bogoych@ed.ac.uk, pinzhen.chen@ed.ac.uk

Abstract

Terminology correctness is important in the downstream application of machine translation, and a prevalent way to ensure this is to inject terminology constraints into a translation system. In our submission to the WMT 2023 terminology translation task, we adopt a translate-then-refine approach which can be domain-independent and requires minimal manual efforts. We annotate random source words with pseudo-terminology translations obtained from word alignment to first train a terminology-aware model. Further, we explore two post-processing methods. First, we use an alignment process to discover whether a terminology constraint has been violated, and if so, we re-decode with the violating word negatively constrained. Alternatively, we leverage a large language model to refine a hypothesis by providing it with terminology constraints. Results show that our terminology-aware model learns to incorporate terminologies effectively, and the large language model refinement process can further improve terminology recall.

1 Introduction

One of the major obstacles encountered by neural machine translation (NMT) systems pertains to the utilization of suitable domain-related words when translating specialized content not present in the training data. An illustrative instance of this challenge arises when translating "transformer" from English into another language, where the accurate translation depends on the context or the preference of the audience (Figure 1). A straightforward literal translation approach often leads to suboptimal outcomes, prompting human translators unfamiliar with domain-specific knowledge to resort to reference materials for terminology precision. This issue is prevalent in the translation industry, with many commercial translation service providers offering paid solutions to address it. Furthermore, it

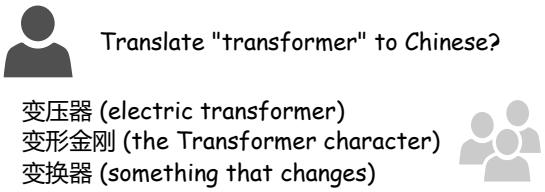


Figure 1: Terminology hints can help disambiguate polysemantic words when translating with limited context.

is a popular area in machine translation research, indicated by efforts such as WMT shared tasks organization and participation focusing on terminology and domain-specific translations (Alam et al., 2021; Bawden et al., 2019, 2020, inter alia).

This year's WMT terminology translation task features three language directions: German-to-English, Chinese-to-English, and English-to-Czech. In addition to reading in a source sentence, participating systems need to employ a provided dictionary, which contains source-target terminology word mappings, to incorporate into the target translation. For each source sentence in the test set, there are three modes of applying terminology constraints:

1. *Terminology* constraint: Dictionaries of real terminology words are provided, to be incorporated in the translations.

2. *Random* constraint: Random (but presumably correct) word mappings are obtained using a word alignment tool and provided as a pseudo-terminology dictionary.

3. *No* constraint: Source sentences can be freely translated without external information.

We interpret that the no-constraint setting allows us to measure the competing systems' quality and understand to what degree the systems effectively utilize the provided random and terminology dictionaries. Our baseline approach is to train

*Equal contribution.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 890–896
December 6–7, 2023.

a terminology-aware translation (TAT) system inspired by Dinu et al. (2019), where, in the training data, source words are tagged with desired translations inline on the source side. Then we propose two separate refinement strategies on top of it to aggressively encourage the appearance of terminologies:

1. We use a neural word aligner to identify terminology constraints missed by the baseline system, and use the same system to re-decode the source by negatively constraining (disallowing) previously incorrectly translated tokens.

2. We also investigate the capability of a large language model to simultaneously paraphrase an existing translation to include the desired terminology constraints via curated prompts.

Our proposed techniques can incorporate target terminology words with around 80% recall, using automatic and soft constraints in a two-step refinement process. We observe that for German-English, our terminology-aware training and negatively constrained decoding perform better, whereas, for Chinese-English and English-Czech, LLM-based refinement achieves higher scores. In terms of overall translation accuracy, we find that negatively constrained decoding could lead to a tiny drop and LLMs are able to maintain or improve quality according to a reference-free neural metric.

2 Related Work

Previous research on terminology translation could be divided into two categories: *soft* constraint and *hard* constraint, depending on whether the resulting translation system will enforce the appearance of desired target translations. In the soft constraint setting, the convention is to train a model that is able to ingest the target terminology words inline, directly placing them after the corresponding source words in the source input (Dinu et al., 2019). Many later implementations stem from this to include new elements such as additional lemmatization (Bergmanis and Pinnis, 2021) or grammatical error correction (Pham et al., 2021) as a post-processing step in order to achieve a more fluent output. Instead of placing the target constraint words inline, some other works train a system that takes the terminology constraint as either a prefix or a suffix (Jon et al., 2021; Turcan et al., 2022).

Most hard constraint work involves post-processing a translation with desired terminologies. Post et al. (2019) inserted untranslatable tokens (also known as placeholders) into the source, which will remain unchanged through the translation process. Then the placeholders are replaced with terminology words in the target language. This is entirely performed as a post-processing step. Such terminology replacement could also be done by keeping and replacing the source word at inference time, and it is also feasible to run target word replacement as post-processing (Molchanov et al., 2021). A hard constraint method guarantees that the chosen terminology token will appear, but often results in less fluent output, especially for morphologically rich languages because the context is not taken into consideration during replacement. It also mandates more complicated post-processing than the soft constraint approaches.

Our first post-processing proposal relies on constrained decoding, which refers to either allowing certain tokens or blocking specific tokens during inference time (Hokamp and Liu, 2017). It has been applied to terminology injection, paraphrasing, parallel sentence mining, etc (Hasler et al., 2018; Kajiwara, 2019; Chen et al., 2020). We opt for negatively constraining the tokens that violated the given terminology alignments by preventing them from entering the hypothesis beam in the refinement stage. These alignments are computed using word alignment tools (Dyer et al., 2013; Dou and Neubig, 2021).

Another post-processing method in our study prompts an LLM to refine a translation and incorporate terminology terms simultaneously. Whilst previous studies have explored the translation capability of LLMs (Vilar et al., 2023; Zhang et al., 2023), the works closely relevant to us are from Moslem et al. (2023) and Ghazvininejad et al. (2023). We adopt the paradigm from the latter, which re-words a constraint dictionary as a natural text and affixes it into a translation prompt. While they focused on rare words without directly benchmarking on terminology translation, our post-processing step can be seen as an extension of word-level controlled prompting to terminology translation with large language models. Both of our post-processing methods should be categorized as soft constraint approaches since there is no guarantee that negatively constrained decoding or an LLM will necessarily incorporate the constraints in a re-generation.

3 Terminology-Aware Training

The goal of our system implementation is to create a general-purpose terminology-aware translation system that is unsupervised and domain-agnostic, and requires the minimum effort of pre- and post-processing.

3.1 Terminology creation

Inspired by Dinu et al. (2019), we applied terminology constraints during training, but a key difference is that, unlike their approach, we assume that we have no access to downstream domain or terminology constraints during training, in order to build a general-purpose domain-agnostic system. Consequently, we have no curated terminology data to use. Therefore, we generate (pseudo-)terminology information using word alignments. Our workflow can be detailed as:

1. We compute the word alignment information for the entire training set using `fast_align` (Dyer et al., 2013).

2. For each sentence, we select all bijective source-target mappings as our terminology candidates. We also filter out trivial mappings where the source and target tokens are the same (e.g. numbers, names), because those mappings are simple and hence likely to be correctly translated by a translation system even without any terminology awareness.

3. In the training data, we replace $\texttt{srcword}_\text{i}$ in the source sentence with:
   $\texttt{srcword}_\text{i}$ `__target__` $\texttt{trgword}_\text{j}$ `__done__`
   where the $srcword_\text{i}$ is the i-th source word inside the sentence, and $trgword_\text{j}$ is the word inside the target sentence, corresponding to srcword_i according to word alignment information. This replacement occurs with around 10% probability for each candidate source-target pair. For a sentence that does not have an associated terminology constraint, the data is the same as normal NMT.

4. At inference time, we process the test data similarly to above, except that the source-target word mapping comes from a supplied terminology dictionary.

In practice, our translation system is trained with a mix of normal translation data and terminology-injected data. The advantage of this strategy is that the trained models are general-purpose, so they can translate normal texts without terminology injection. Further, they have been exposed to a wide variety of constraints during training, making them robust to potentially unseen domain constraints.

Overall, our method is very similar to Bergmanis and Pinnis (2021)'s work, except that we use whole words but not lemmas to ease pre-processing. We presume that the language model will be able to adjust the terminologies accordingly, especially for morphologically rich languages on the target side. This enables our method to be trivially transferable across languages.

Finally, our systems could easily be turned into hard-constrained by replacing the source word with the desired target terminology word. This could be feasible because our terminology-aware training installs the copying behaviour in the neural translation model, although in this mode the model would produce markedly less fluent output.

3.2 Model architecture

We trained Transformer-style machine translation models (Vaswani et al., 2017) using the Marian NMT toolkit (Junczys-Dowmunt et al., 2018). We used the `Transformer-Big` preset which is a 6 encoder, 6 decoder architecture with 1024 hidden size, and 4096 feedforward size.[1]

3.3 Data

The terminology task uses the same data as the constrained condition in the WMT23 general translation task. We carefully cleaned, filtered, and deduplicated the available WMT training sets provided by the organisers, as well as the available back-translation data. After preprocessing we were left with the following:

- German-to-English (`de-en`): 199M lines of parallel data and 29.5M lines of back-translated data.

- Chinese-to-English (`zh-en`): 21.8M lines of parallel data and 15.6M lines of back-translated data.

- Czech-to-English (`cs-en`): 61.8M lines of parallel data and 57M lines of back-translated data.

[1] https://github.com/marian-nmt/marian/blob/master/src/common/aliases.cpp#L114

Query	Prompt template
Translation	`Source: ${source}` `Please give me a translation in ${lang} without any explanation.`

Refinement	`Source: ${source}`
	`Translation: ${translation}`
	`Please give me a better ${lang} translation without any explanation.`
	`"${srcword$_0$}" should be translated as "${trgword$_0$}";`
	`"${srcword$_1$}" should be translated as "${trgword$_1$}";`
	`...`
	`"${srcword$_k$}" should be translated as "${trgword$_k$}".` *(with k >= 0)*

Table 1: Large language model prompt templates for unconstrained and constrained translation.

3.4 General quality

The quality of our models without terminology translation is shown in Table 2, where we report BLEU (Papineni et al., 2002) and COMET_{DA}[2] (Rei et al., 2020) scores on test sets from the WMT22 general translation task. We note that terminology augmentation during training could result in a slight quality drop.

	BLEU	COMET_{DA}
de-en	31.3	0.8334
en-cs	39.5	0.8715
zh-en	20.3	0.7559

Table 2: Performance of our terminology-aware translation systems in the WMT22 general translation task.

4 Post-Translation Terminology Injection

Despite training our model with terminology awareness, there is no mechanism to ensure that the desired terminology constraint will appear on the target side. The neural network decoding behaviour is not entirely predictable, especially given the assumption of no additional domain adaptation. Below, we present two distinct strategies to try *harder* to promote the terminology constraints, via automatic post-editing through constrained beam search and large language models.

4.1 Negatively constrained decoding

While it is easy enough to notice when a target terminology term is not generated as per a given constraint, it is not trivial to understand which word has been produced in place of the desired term. In order to do this, we make use of *awesome-align*, a neural multilingual word aligner (Dou and Neubig, 2021), with the following procedure:

1. For each source-translation pair, we check if all required terminology terms appear on the target side. If they do, then we stop processing more rules.

2. Then, we use *awesome-align* to compute word alignments and detect the word(s) that have been generated in place of the desired terms according to the provided terminology constraints.

3. We decode the source sentence again, penalising the words that violated the terminology constraint, by forbidding the decoder from generating them at each generation step, unless they carry more than 95% of the probability mass at a certain step.

In practice, this procedure can be repeated infinitely, until all terminology constraints are fulfilled, but we decided to limit it to only one iteration, to keep this a realistic production scenario in terms of computational budget.

4.2 Large language models

Recent years saw the rise of large language models (LLMs), which have a strong capability in various NLP tasks. In this paper, we investigate the effectiveness of using a large language model to generate terminology terms during translation by adding constraints to Chen et al. (2023)'s translation refinement prompts. We use two distinct prompts: free translation and translation refinement queries. The translation query sends a source sentence and

[2] `wmt22-comet-da`. This is a reference-based metric which requires the source input, hypothesis, and reference.

Mode	Model	Refine	de→en		zh→en		en→cs	
			Recall	COMET_{QE}	Recall	COMET_{QE}	Recall	COMET_{QE}
terminology constraints	TAT	-	**82.30**	.0797	49.98	-.0896	73.75	.0601
	TAT	NCD	82.01	.0775	50.42	-.0903	73.26	.0588
	TAT	LLM	64.35	.1197	**83.06**	.0185	76.00	.0866
	LLM	-	41.86	.1244	46.63	.0191	48.14	.0913
	LLM	LLM	70.48	.1180	81.01	.0201	**78.94**	.0882
no constraint†	TAT	-	39.82	.1085	13.64	-.1163	48.11	.0712
	TAT	LLM	39.59	.1251	42.76	.0203	47.31	**.0955**
	LLM	-	41.86	.1244	46.63	.0191	48.14	.0913
	LLM	LLM	39.65	**.1258**	46.72	**.0228**	46.22	.0943
random constraints	TAT	-	**76.17**	.0716	81.55	-.1105	57.10	.0502
	TAT	NCD	75.79	.0698	**82.03**	-.1123	56.42	.0465
	TAT	LLM	61.46	.1206	63.17	.0175	70.97	.0875
	LLM	-	38.70	.1244	52.49	.0191	39.34	.0913
	LLM	LLM	66.74	.1188	67.10	.0196	**73.37**	.0867
no constraint‡	TAT	-	35.60	.1085	36.18	-.1163	37.35	.0712
	TAT	LLM	37.58	.1251	49.48	.0203	39.03	**.0955**
	LLM	-	38.70	.1244	52.49	.0191	39.34	.0913
	LLM	LLM	37.62	**.1258**	49.00	**.0228**	38.42	.0943

†Recall computed against terminology constraints.

‡Recall computed against random constraints.

Table 3: Terminology recall and translation quality measured by COMET_{QE} of our systems on the *blind test* set. TAT: terminology-aware translation; NCD: negatively constrained decoding; LLM: large language model.

requests a translation in the target language without any other information. On the other hand, the refinement query feeds back an unconstrained translation together with terminology constraints to request a new translation. This essentially forms an LLM version of the constrained beam search discussed in Section 4.1. The constraints are enforced through natural language instructions in the prompts, under the situation where the softmax distribution from an LLM is not accessible by users.

The LLM we use is OpenAI's GPT-3.5.[3] It is a closed-source commercial system, where the model weights and the inference states are not available to users. The model has a context window of 4096 which is sufficient to cover an instruction, a source sentence, several terminology constraints, as well as the target translation. It is public to all users at a relatively cheap cost. In our settings, each translation is carried out in a new query session.

In Table 1 we outline the two prompt templates we used. During querying, the placeholder variables are substituted with corresponding string values. For the refinement query, when a terminology dictionary is supplied, the source and target words are fed to the LLM via the prompt (Ghazvininejad et al., 2023); if there is no terminology dictionary, the query simply asks for a refined translation. The two-step experiment with LLMs can be summarized as follows:

1. We obtain an initial unconstrained translation, which may or may not fulfil all the terminology constraints. It can come from either the LLM itself or the terminology-aware translation model built in Section 3.1.

2. We query the LLM with the constrained translation prompt to obtain a refined translation with terminology incorporated in the prompt.

5 Results and Discussions

We present our *blind test* results in Table 3, which include both terminology recall and COMET_{QE} scores computed by us.[4] We used COMET_{QE} in particular because it does not require references

[3] `gpt-3.5-turbo-0613`, a snapshot of the GPT-3.5 model on 13 June 2023

[4] `wmt21-comet-da-qe`

which are not accessible to us. We assess the effectiveness of our methods by comparing the terminology recall of our systems with and without applying terminology constraints, in both *random* and *real terminology* scenarios.

5.1 Translation quality

In terms of translation quality reflected in COMET_{QE}, we observe that the LLM rows attain superior results, which is not surprising considering that we use an unconstrained commercial model GPT-3.5. By comparing TAT with TAT+NCD, or comparing LLM with LLM+LLM under a constrained scenario, we conclude that applying terminology constraints usually lead to a sacrifice in translation quality regardless of the language direction or the systems involved. Nonetheless, as a contrasting experiment with no constraint, LLM+LLM achieves a slightly better COMET_{QE} score than using an LLM to translate without refinement.

Our model performed poorly on the `zh-en` task in terms of COMET_{QE} scores. We suspect that this is because of the domain mismatch between the translation data from the general domain and the Chinese terminology test set. Upon manual inspection, we found that the latter includes web novels and literal writing which are likely to be under-represented in the generic training data.

5.2 Terminology recall

Focusing on terminology generation, compared with TAT or LLM in unconstrained settings, TAT marks 30-40 higher recall of terminology terms in the constrained *terminology* and *random* settings. This indicates that our terminology-aware training is effective in teaching translation models to follow customized source-target word alignments.

Next, as a post-processing step, negatively constrained decoding seems to be disappointing in practice. TAT+NCD often produces worse results than TAT alone in terms of both quality and terminology recall, except for `zh-en` with *random* constraints. We hypothesize that this could be due to two problems: (1) word alignment errors could propagate into this process, and (2) by applying NCD, we might capture a missed terminology term but at the cost of mis-translating other words. Our constraining procedure might be improved by performing shortlisting, namely positively constrained decoding, as opposed to negatively limiting the beam search in an iterative approach.

We find the results promising when using LLMs for terminology injection. Looking at LLM+LLM versus LLM alone in various constrained conditions, terminology recall improves significantly with very little drop in overall quality. Also by comparing TAT+LLM with TAT alone, we observe that TAT and LLMs each have their own merits depending on the language direction. In terms of recall, TAT wins in `de-en`, TAT+LLM wins in `zh-en`, and they are close in `en-cs`. However, TAT+LLM is way ahead if measured by COMET_{QE}. However, we must note that an LLM costs significantly more resources than a dedicated translation model at both training and inference time.

6 Conclusion and Future Work

We participated in all tracks of the WMT 2023 terminology shared task with a terminology-aware translation baseline, and two distinct refinement procedures using negatively constrained beam search and large language models separately. The results we produced gave us insights into the pros and cons of our systems. In future work, we could explicitly enforce the generation of the terminology token by identifying the appropriate time step and manipulating the probability distribution after softmax computation, even in an open-source large language model. This is not entirely trivial due to the presence of subwords but could be achievable.

Acknowledgement

This project has received funding from UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant numbers 10052546 and 10039436].

Some computations described in this research were performed using the Baskerville Tier 2 HPC service (https://www.baskerville.ac.uk/). Baskerville was funded by the EPSRC and UKRI through the World Class Labs scheme (EP/T022221/1) and the Digital Research Infrastructure programme (EP/W032244/1) and is operated by Advanced Research Computing at the University of Birmingham.

References

Md Mahfuz Ibn Alam, Ivana Kvapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Kweonwoo Jung, Philipp Koehn, and Vassilina Nikoulina. 2021. Find-

ings of the WMT shared task on machine translation using terminologies. In *Proceedings of WMT*.

Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurelie Neveol, Mariana Neves, Felipe Soares, Amy Siu, Karin Verspoor, and Maika Vicente Navarro. 2019. Findings of the WMT 2019 biomedical translation shared task: Evaluation for MEDLINE abstracts and biomedical terminologies. In *Proceedings of WMT*.

Rachel Bawden, Giorgio Maria Di Nunzio, Cristian Grozea, Inigo Jauregi Unanue, Antonio Jimeno Yepes, Nancy Mah, David Martinez, Aurélie Névéol, Mariana Neves, Maite Oronoz, Olatz Perez-de Viñaspre, Massimo Piccardi, Roland Roller, Amy Siu, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Dina Wiemann, and Lana Yeganova. 2020. Findings of the WMT 2020 biomedical translation shared task: Basque, Italian and Russian as new additional languages. In *Proceedings of WMT*.

Toms Bergmanis and Mārcis Pinnis. 2021. Facilitating terminology translation with target lemma annotations. In *Proceedings of EACL*.

Pinzhen Chen, Nikolay Bogoychev, Kenneth Heafield, and Faheem Kirefu. 2020. Parallel sentence mining by constrained decoding. In *Proceedings of ACL*.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2023. Iterative translation refinement with large language models. *arXiv preprint*.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of ACL*.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of EACL*.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of NAACL-HLT*.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *arXiv preprint*.

Eva Hasler, Adrià de Gispert, Gonzalo Iglesias, and Bill Byrne. 2018. Neural machine translation decoding with terminology constraints. In *Proceedings of NAACL-HLT*.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of ACL*.

Josef Jon, Michal Novák, João Paulo Aires, Dusan Varis, and Ondřej Bojar. 2021. CUNI systems for WMT21: Terminology translation shared task. In *Proceedings of WMT*.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL*.

Tomoyuki Kajiwara. 2019. Negative lexically constrained decoding for paraphrase generation. In *Proceedings of ACL*.

Alexander Molchanov, Vladislav Kovalenko, and Fedor Bykov. 2021. PROMT systems for WMT21 terminology translation task. In *Proceedings of WMT*.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of EAMT*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*.

Minh Quang Pham, Josep Crego, Antoine Senellart, Dan Berrebbi, and Jean Senellart. 2021. SYSTRAN @ WMT 2021: Terminology task. In *Proceedings of WMT*.

Matt Post, Shuoyang Ding, Marianna Martindale, and Winston Wu. 2019. An exploration of placeholding in neural machine translation. In *Proceedings of MT Summit*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of EMNLP*.

Elsbeth Turcan, David Wan, Faisal Ladhak, Petra Galuscakova, Sukanta Sen, Svetlana Tchistiakova, Weijia Xu, Marine Carpuat, Kenneth Heafield, Douglas Oard, and Kathleen McKeown. 2022. Constrained regeneration for cross-lingual query-focused extractive summarization. In *Proceedings of COLING*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of ACL*.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *Proceedings of ICML*.

Lingua Custodia's participation at the WMT 2023 Terminology shared task

Jingshu Liu **Mariam Nakhlé** **Gaëtan Caillaut** **Raheel Qader**
Lingua Custodia, France
{jingshu.liu,mariam.nakhle,gaetan.caillaut,raheel.qader}@linguacustodia.com

Abstract

This paper presents Lingua Custodia's submission to the WMT23 shared task on Terminology shared task. Ensuring precise translation of technical terms plays a pivotal role in gauging the final quality of machine translation results. Our goal is to follow the terminology constraint while applying the machine translation system. Inspired by the recent work of terminology control, we propose to annotate the machine learning training data by leveraging a synthetic dictionary extracted in a fully non supervised way from the give parallel corpora. The model learned with this training data can then be then used to translate text with a given terminology in a flexible manner. In addition, we introduce a careful annotated data re-sampling step in order to guide the model to see different terminology types enough times. In this task we consider all the three language directions: Chinese to English, English to Czech and German to English. Our automatic evaluation metrics with the submitted systems show the effectiveness of the proposed method.

1 Introduction

It is well proven that modern Neural Machine Translation (NMT) systems (Sutskever et al., 2014; Bahdanau et al., 2014; Luong et al., 2015; Vaswani et al., 2017) achieve generally satisfying translation results. Nonetheless, the performance of translation with terminology control remains to be improved. This paper describes our submission to the WMT23 Terminology translation task in Chinese to English, English to Czech and German to English direction. The task aims to develop and evaluate machine translation systems which can translate domain specific terms in an accurate and consistent way with some extra terminology information. Note that the terminology is provided only in the inference phase, for the training there's no existing resources about the terminology.

Previous works on machine translation with terminology control can be grouped into two categories according to whether the method needs training the model with terminology information. One group incorporates the constraint during the inference time(Hokamp and Liu, 2017; Post and Vilar, 2018; Susanto et al., 2020). These methods can typically satisfy most of the constraints but suffer from high computational cost and sometimes low translation quality because it always tries to strictly apply the terminology constraint regardless of the correctness of the whole sentence. The other group integrates lexical constraints during training (Dinu et al., 2019; Crego et al., 2016; Song et al., 2019) by annotating the data with special tags in order to guide the model to learn the enforcement of the translation constraints. The main disadvantage of these methods is the lack of guarantee of all constraints in the translations. Another limitation of these works is that they usually requires a term dictionary to augment the data, such extra resource is not always trivial to obtain for some domains in some languages.

Our work follows the second line of methods which incorporate terminology in the training by inserting special tags. Upon the recent works of (Dinu et al., 2019; Ailem et al., 2021), our system has made several improvements:

1. a terminology extracted from the given training corpus in a full non supervised manner rather than from a supervised approach or a given dictionary like in (Ailem et al., 2021).
2. only use special tags without source factors(Dinu et al., 2019) to annotate source and target terms in parallel sentences.
3. use a careful tag sentence re-sampling process to represent various constraint scenarios.

We evaluate our work on all the WMT23 terminology task including the blind test. Since the

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 897–901
December 6–7, 2023.

reference is not released by the time of writing this paper, we evaluate our system by a simple naive strict match with respect to the target constraint. Our results show the effectiveness of the proposed method.

2 Method

In this section we present out system for the terminology task. Our approach is inspired by Ailem et al. (2021) and is further developed and adapted to this task.

2.1 Non supervised bilingual dictionary extraction

Approaches incorporating the constraints during the training time require some pre-built terminologies or dictionaries such as in the terminology task of *WMT2021*. The idea is to create training samples to guide the model to integrate the constraints when generating the output. However, in this year's shared task, terminology is not provided. Previous approaches such as Hazem and Morin (2016) and Liu et al. (2018) require heavy computation. Artetxe et al. (2016) can only learn single word bilingual lexicon. In this work, since our goal is to annotate the training data, having some noise in the extracted dictionary is affordable but the the number of the dictionary entries should be high enough to cover as much as possible different terminology constraint scenarios. Thereby we propose a simple yet efficient non supervised bilingual dictionary extraction approach which yields a large amount of aligned single and multi word items.

Our approach consists in extracting entries from two aspects: first we extract exact matching ngrams which contains more than 50% of non stop word or punctuation tokens from the two language texts, to prevent this process from being unnecessarily long, we limit the ngrams to five; the second aspect consists in extracting a whole sequence which is entirely included in another sequence of the corpus. The final dictionary contains both invariable and long sequence entries.

2.2 Data annotation

Following the work of Dinu et al. (2019) and Ailem et al. (2021), sentences matching source and target contraint terms are annotated with some special tags as illustrated in Figure 1.

Note that we also use mask tokens for the source term since this this provides a more general pattern

Source	His critics state that this will just increase the **budgetary deficit** .
Constraint	**budgetary deficit** → **Haushaltsdefizit**
term annotation	His critics state that this will just increase the <S> **budgetary deficit** <C> **Haushaltsdefizit** </C> .
+MASK	His critics state that this will just increase the <S> **MASK MASK** <C> **Haushaltsdefizit** </C> .

Figure 1: Training data annotation.

for the model to learn to perform the copy operation every time it encounters the tag **<S>** followed by the **MASK** token. Moreover, this makes the model more apt to support conflicting constraints, i.e., constraints sharing the same source part but which have different target parts. This may be useful if some tokens must be translated into different targets for some specific documents and contexts at test time. Our preliminary experiments have shown the effectiveness of adding masks after data annotation.

2.3 Annotated data resampling

After the automatic data annotation, several filters are applied to construct a final tagged data set which equals to 20% of the original data. The goal is to cover different constraint contexts so that the model can learn all possible cases. The criterions of the filters are as follows:

Constraint length. Oversample constraints with more composing tokens.

Constraint occurrence. Oversample constraints with low occurrence.

Constraint number. Oversample sentences with different constraint numbers.

Constraint position. Make sure that constraints at the beginning, middle and end of a sentence follow a distribution of 10%, 80% and 10%.

For all the oversampling, we apply a modified version of the temperature sampling with a temperature equal to 5:

$$P_t s(t) = \frac{P(t)^{1/T}}{\sum_i P(t_i^{1/T})}$$

where $P_t st$ is the temperature sampling probability for term t. T is the hyper-parameter temperature. $P(t)$ is the probability of term t, we assume it can be calculated by the following:

$$P_(t) = \frac{N(t)}{\sum_i N(t_i)}$$

where $N(t)$ is the frequency of term t in the training corpus. So $\sum_i N(t_i)$ represents actually the sum frequency of all terms. Finally the oversample size for term t, $N_{oversample}(t)$ will be the rounded up value of:

$$N_{oversample}(t) = P_ts(t) * \frac{N(t_{max})}{P_ts(t_{max})}$$

where t_{max} is the term having the highest frequency.

3 Experiments

3.1 Data

We participate in all three language pairs: Chinese to English (noted as zh2en), English to Czech (noted as en2cs) and German to English (noted as de2en). We use the corresponding parallel data provided by the general translation task and the development data of the terminology task. Since the given development set has only 100 sentences, we first oversample these 100 sentences by 10 times, then we randomly take 4000 sentences from given general data and add them to the oversampled data. This results in a final development set of 5000 sentences.

Regarding the training data annotation dictionaries, we extract invariable ngrams from one million random sentences. In addition, we follow what we have described in 2.1: sentences which are included in other longer sequences are added to the dictionary. An overview of the data is shown in 1.

Data	**size(sentence/item)**
zh2en train	33 892 215
zh2en dictionary	445 727
en2cs train	130 023 715
en2cs dictionary	559 063
de2en train	288 591 578
de2en dictionary	769 915

Table 1: Data used in the task

3.2 Settings

For all our translation models, we use a Transformer (Vaswani et al., 2017) with 6 stacked encoders/decoders and 8 attention heads as a building block for our systems. We also tie the source and target embeddings with the softmax layer with a shared source and target vocabulary. The model size is 512 for the source and target embeddings, 2048 for the inner layers of the fully connected feed-forward network and a dropout rate of 0.15.

Training batch size is set to 4000 tokens per iteration and we evaluate the model on the development set for every 5 000 iterations. The model is trained with an initial learning rate of 10^{-5} and 10 000 warm up steps. Training stop condition is 15 consecutive checkpoints without improvement. We use a length penalty of 0.65 and a beam size of 5 during inference for all models. All models are trained on two NVIDIA Geforce 2080Ti.

Before annotating the training data, we apply Moses tokenizer (Koehn et al., 2007)and we train a truecaser for each language and then truecase each language pair data. We also use *subword nmt*[1] to train a BPE (Sennrich et al., 2016) model of 50k merges.

3.3 Results

We evaluate our systems on the translation constraint success rate by a simple strict match because by the time of our naive evaluation, the reference was not available. We report our results on the test set in Table 2 with two settings: with and without terminology control.

	Accuracy% †	Accuracy% ‡
German to English	92.59	69.29
English to Czech	94.15	47.43
Chinese to English	83.77	22.21

†: with terminology, ‡: without terminology applied

Table 2: Term strict match accuracy (%) on the WMT23 testset with and without using extra terminology.

As shown in Table 2, our system achieves more than 90% accuracy on German to English and English to Czech test set. While the accuracy is obviously lower (roughly 10 points lower) on the Chinese to English test set, we think this might be related to the higher difficulty of the Chinese to English test set. In the test set, there are some

[1]https://github.com/rsennrich/subword-nmt

constraints which are basically named entity transcribed in *Pinyin*[2] script. For example, 段凌天 ⟶ *Duan Ling Tian* (Person), 武宗学府 ⟶ *the martial arts training institute* (Association). The model needs to somehow learn the transcription from Chinese character to *Pinyin* or a specific alignment on which there aren't much train data in the provided parallel corpus. As a whole, our system shows satisfying results when the terminology is provided. To study whether the high accuracy results are obtained by our terminology control system or not, we also evaluate our system but without giving any terminology during the inference. We should expect a big gap between the two settings (with and without terminology during the inference). The results confirm our assumption: an average of 40+ points of difference for the three directions.

For the blind test, we present our strict match accuracy in Table 3. The data in the blind test is provided in three different modes: the first one corresponds to general machine translation and the second one has the terminology dictionary added. The data is provided in three different modes. The last one has random, though correct, translations of words, which are not terminologies. The idea is to see if we obtain improvement between the different modes, in which case it means that the model is good at terminology control not because that it has learned the specific way of translating those terms but has learned how to make good use of terminology information.

	Accuracy% †	Accuracy% §	Accuracy% ‡
German to English	97.35	98.18	36.16
English to Czech	94.76	94.50	45.06
Chinese to English	93.26	74.20	48.45

†: with correct terminology; §: with random terminology; ‡: without terminology applied

Table 3: Term strict match accuracy (%) on the WMT23 blind testset with correct and random term, and without using extra terminology.

On all the language directions, our system achieve more than 90% accuracy when the terminology information is provided. When a random constraint is given, we consider the given random constraint term as the reference translation. In this case, we observe that our system can still obtain a high accuracy score. This means that the model is able to generalize the behavior of outputting any constraint. Finally, in the general translation setting, we see a sharp decreasing of the accuracy, 40+ points lower compared to the terminology setting. This phenomenon shows that the model is not just good on its own but can make good use of the terminology.

[2]en.wikipedia.org/wiki/Pinyin

4 conclusion

This paper describes our submission to the terminology shared task. We participate in three language directions, German to English, English to Czech and Chinese to English. We extract a bilingual dictionary for the three language directions in a fully non supervised way and train a neural machine translation model with augmented data for each direction. Our term strict match evaluation shows the effectiveness our proposed system for all the three directions.

5 Limitations

Since we pursue the line of works which incorporate terminology control by adding special tags during the training. This system has also the limit of not being able to guarantee the constraints to be present in the output because of the soft nature. This is mainly concertized by two cases:

- **No constraint**. The constraint is not presented at all in the translation.
- **Variant constraint**. The exact format of the constraint is not presented but a variant is proposed in the output.

We observe that for most of the time when the model fails to generate the target constraint, the scenario belongs to the second case which proposes a variant of the constraint. This translation is acceptable in a human evaluation context from time to time.

To address this main limit, we would like to exploit assembling our method with other techniques such as a post processing step to force the constraint if the constraint is not presented in the output.

References

Melissa Ailem, Jingshu Liu, and Raheel Qader. 2021. Encouraging neural machine translation to satisfy terminology constraints. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1450–1455, Online. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word

embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP'16)*, pages 2289–2294, Austin, TX, USA.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Josep Crego, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart, Egor Akhanov, Patrice Brunelle, Aurelien Coquard, Yongchao Deng, et al. 2016. Systran's pure neural machine translation systems. *arXiv preprint arXiv:1610.05540*.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, page 3063–3068.

Amir Hazem and Emmanuel Morin. 2016. Efficient data selection for bilingual terminology extraction from comparable corpora. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING'16)*, pages 3401–3411, Osaka, Japan.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, page 1535–1546.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Jingshu Liu, Emmanuel Morin, and Sebastián Peña Saldarriaga. 2018. Towards a unified framework for bilingual terminology extraction of single-word and multi-word terms. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING'18)*, pages 2855–2866, Santa Fe, NM, USA.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, page 1412–1421.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. *Proceedings of NAACL-HLT 2018*, page 1314–1324.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019. Code-switching for enhancing nmt with pre-specified translation. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, page 449–459.

Raymond Hendy Susanto, Shamil Chollampatt, and Liling Tan. 2020. Lexically constrained neural machine translation with levenshtein transformer. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 3536–3543.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Domain Terminology Integration into Machine Translation: Leveraging Large Language Models

Yasmin Moslem[1,2,§], Gianfranco Romani[3], Mahdi Molaei[4], Rejwanul Haque[1,5], John D. Kelleher[1,6], and Andy Way[1,2]

[1]ADAPT Centre

[2]School of Computing, Dublin City University, Dublin, Ireland

[3]Thomson Reuters, Zug, Switzerland

[4]Department of Computer Engineering, University of Tabriz, Tabriz, Iran

[5]Department of Computing, South East Technological University, Carlow, Ireland

[6]Hamilton Institute, Maynooth University, Maynooth, Ireland

Abstract

This paper discusses the methods that we used for our submissions to the WMT 2023 Terminology Shared Task for German-to-English (DE-EN), English-to-Czech (EN-CS), and Chinese-to-English (ZH-EN) language pairs. The task aims to advance machine translation (MT) by challenging participants to develop systems that accurately translate technical terms, ultimately enhancing communication and understanding in specialised domains. To this end, we conduct experiments that utilise large language models (LLMs) for two purposes: generating synthetic bilingual terminology-based data, and post-editing translations generated by an MT model through incorporating pre-approved terms. Our system employs a four-step process: (i) using an LLM to generate bilingual synthetic data based on the provided terminology, (ii) fine-tuning a generic encoder-decoder MT model, with a mix of the terminology-based synthetic data generated in the first step and a randomly sampled portion of the original generic training data, (iii) generating translations with the fine-tuned MT model, and (iv) finally, leveraging an LLM for terminology-constrained automatic post-editing of the translations that do not include the required terms. The results demonstrate the effectiveness of our proposed approach in improving the integration of pre-approved terms into translations. The number of terms incorporated into the translations of the blind dataset increases from an average of 36.67% with the generic model to an average of 72.88% by the end of the process. In other words, successful utilisation of terms nearly doubles across the three language pairs.

1 Introduction

The primary goal of the WMT 2023 Terminology Shared Task is to evaluate the ability of MT systems to accurately translate technical terminology. The task aims to assess the extent to which MT models can utilise additional information regarding the translation of terminology. The shared task requires the participants to provide three translations, one without terms and the others with two individual sets of terms.

There have been several advancements in the area of MT domain adaptation, where an MT model is expected to follow the style and terminology of a certain domain or client (Chu et al., 2017; Kobus et al., 2017). Moreover, some researchers give special focus to terminology while training and fine-tuning MT systems (Dinu et al., 2019; Hu et al., 2019b; Haque et al., 2020; Michon et al., 2020; Nayak et al., 2023). However, forcing an MT model to adhere to certain terminology at inference time is among the most challenging aspects of MT. Hence, several researchers have investigated approaches to terminology-constrained decoding at translation time (Hokamp and Liu, 2017; Hasler et al., 2018; Post and Vilar, 2018; Hu et al., 2019a; Exel et al., 2020). The goal is to ensure that the MT system can accommodate unseen terminology while retaining translation accuracy and fluency.

Recently, since the emergence of advanced LLMs such as GPT-3 (Brown et al., 2020), BLOOM (Le Scao et al., 2022), PaLM (Chowdhery et al., 2022), Falcon (Penedo et al., 2023), Llama 2 (Touvron et al., 2023), and Jais (Sengupta et al., 2023) to mention just a few, researchers have been exploring the capabilities of these models for a number of tasks including MT (Bawden and Yvon, 2023; Hendy et al., 2023; Jiao et al., 2023; Moslem et al., 2023; Vilar et al., 2023). Some work investigates whether it is possible to utilise LLMs for terminology-constrained MT using a pre-defined glossary (Moslem et al., 2023) or even a dictionary (Ghazvininejad et al., 2023). They found the approach is generally effective in increasing the number of terms used in the translation, even for low-resource languages.

§Correspondence: first_name.last_name@adaptcentre.ie

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 902–911
December 6–7, 2023.

We highlight our key contributions with the systems that we submitted for the WMT 2023 Terminology Shared Task as follows:

- **LLMs for domain-specific data augmentation:** In our previous work (Moslem et al., 2022), we employed LLMs, namely GPT-J (Wang and Komatsuzaki, 2021) and mGPT (Shliazhko et al., 2022), to generate domain-specific datasets based on the target sentences in a small authentic dataset, then generated the source sentences with back-translation (Sennrich et al., 2016; Poncelas et al., 2019), and finally fine-tuned an encoder-decoder MT model on this data. In this work, we take a couple of steps forward by instructing an LLM, namely ChatGPT (Brown et al., 2020; Ouyang et al., 2022), to generate terminology-based bilingual synthetic data. In other words, the LLM will generate both the source and target sides of translation pairs, making sure the pre-approved target terms provided by the organisers are used in the translations.

- **LLMs for terminology-constrained MT and MT post-editing:** In our previous work, we utilised an LLM for translation and provided it with a list of terms to support in-context learning, which improved adherence to the required terminology at inference time (Moslem et al., 2023). We also investigated whether we could use an LLM for post-editing MT generated by other systems. In this work, we prompt ChatGPT to insert missing terms into translations generated by an encoder-decoder MT system. In other words, if some of the translations generated by a fine-tuned MT model still do not include the terms provided by the organisers, we feed these translations into an LLM, namely ChatGPT, instructing it to incorporate these terms while using the same translation.

2 Method

In our submissions to the WMT 2023 Terminology Shared Task, we followed these steps:

(i) Generate bilingual synthetic data based on the pre-approved terms, using an LLM, namely ChatGPT.

(ii) Fine-tune a generic model, OPUS (Tiedemann and Thottingal, 2020), on a mix of the terminology-based synthetic data generated in (i) and a randomly sampled portion of the original generic training data.

(iii) Generate translations of the dev, test, and blind datasets provided by the organisers with the fine-tuned model from (ii).

(iv) Apply terminology-constrained automatic post-editing using ChatGPT to incorporate missing terms into translations that do not yet include the required terminology.

2.1 Synthetic Data Generation

We used ChatGPT "gpt-3.5-turbo"[2] to generate bilingual sentence pairs, using the terms provided by the organisers. So, given a target term, the model was asked to generate multiple translation pairs, including both the source (e.g. German) and the target (e.g. English). For parameters of ChatGPT's API, we used *top_p* 1 and *temperature* values 0 and 0.3 to generate diverse outputs.

Example prompt: Terminology-based generation

Please use the "Federal Ministry of Science" to generate just 20 numbered sentences in German-English in one Python dictionary format.

To filter the generated data, we first removed duplicate sentences from the whole dataset, based on both the source and target. Then, we applied language detection of both sides of the data using *fastText*[3] and *pycld2*[4] libraries to ensure that the generated sentences were in our desired languages. We excluded any sentences whose scores were below a certain threshold, namely 0.9 for *fastText* and 90 for *pycld2*.

The filtering step removed less than 1% of the generated data. However, due to computational resource and time limitations, we could not use all the generated data. Table 1 reports the number of generated, filtered, and used translation pairs.

Initially, we only had the development and test datasets, so we used them for the German-to-English language pair. Later, when the organisers released the blind dataset, we used the development, test and blind datasets for the Chinese-to-English and English-to-Czech language pairs.

[2]The model "gpt-3.5-turbo" is a relatively efficient and cost-effective option, so we wanted to understand the quality we can achieve with it.

[3]https://fasttext.cc/docs/en/language-identification.html

[4]https://github.com/aboSamoor/pycld2

Lang	Raw	Filtered	Used
DE-EN	124,215	104,318	68,265
EN-CS	187,471	103,797	64,218
ZH-EN	90,538	72,695	49,001

Table 1: Terminology-based bilingual data generated by ChatGPT for fine-tuning the OPUS model

To assess the quality of the bilingual data generated by ChatGPT, we computed cross-entropy scores (Moore and Lewis, 2010) of the synthetic translation pairs based on the strong encoder-decoder MT model, NLLB-200 3.3B (Costa-jussà et al., 2022). For scoring, we used CTranslate2[5] (Klein et al., 2020) *score_batch()* method with the parameters *batch_type* "tokens" and *max_batch_size* 2024. We scored each synthetic translation pair generated by ChatGPT, and then calculated the average score for the whole dataset. Computing dual cross-entropy scores according to two inverse translation models trained on clean data is an effective method to evaluate data quality (Junczys-Dowmunt, 2018). Hence, we computed the scores of both directions of each language pair according to the multilingual MT model NLLB-200 3.3B because both directions are generated by ChatGPT. To produce a baseline for translation quality, we generated the translations of the same datasets using NLLB-200 3.3B for each language direction with *beam_size* 4, and then scored these translations with the same model. As the scores are in the form of negative log probabilities, we converted them to their exponential equivalents for readability, which are reported in Table 2. It is normal that the model NLLB-200 generates higher scores for its own translations; however, we wanted to know to what extent such scores are comparable to those of ChatGPT's synthetic translation pairs. According to the scores, the German↔English language pair had the most comparable quality, followed by Czech↔English, and Chinese ↔English language pairs.

Among the approaches that can be employed for assessing the quality of synthetic bilingual data is semantic similarity between the two sides of each translation pair (e.g. with mUSE (Yang et al., 2020)). However, the scoring approach that we previously described and used achieves a similar goal while comparing the quality of the synthetic bilingual data to the translation quality of a strong MT baseline model, namely NLLB-200 3.3B.

[5] https://github.com/OpenNMT/CTranslate2

Lang	ChatGPT	NLLB	Diff.
DE-EN	0.59	0.68	0.09
EN-DE	0.56	0.64	0.08
Avg.	0.58	0.66	0.08
CS-EN	0.58	0.70	0.12
EN-CS	0.49	0.58	0.09
Avg.	0.54	0.64	0.10
ZH-EN	0.39	0.56	0.17
EN-ZH	0.09	0.34	0.25
Avg.	0.24	0.45	0.21

Table 2: Scores of translation pairs generated by ChatGPT based on the NLLB-200 3.3B model

2.2 Fine-tuning

Using the term-based synthetic bilingual data generated in the previous step, we fine-tuned encoder-decoder Transformer-based MT models (Vaswani et al., 2017). In particular, we fine-tuned OPUS MT models, with Hugging Face Transformers.[6] We applied mixed fine-tuning (Chu et al., 2017); in other words, we fine-tuned the baseline model with a mix of the terminology-based synthetic data generated from the previous step (cf. Section 2.1) and a randomly sampled portion of the original generic data used to train the OPUS baseline model. The numbers of segments taken from the OPUS generic data are as follows: CS: 372,928, DE: 419,881, ZH: 462,780. We over-sampled the synthetic terminology-based data to make it the same size as the used portion of generic data. The fine-tuning parameters are as follows: *train* = 0.9, *val* = 0.1, *batch_size* = 32, *learning_rate* = 2e-5, *accumulate_gradient* = 4, *weight_decay* = 0.01, *num_train_epochs* = 1, *max_input_length* = 256, *max_target_length* = 256. Finally, we used the fine-tuned model to generate translations for the development, test, and blind sets.

At first glance, the fine-tuning step might look redundant if the LLM can achieve the same translation quality directly, either via zero-shot translation or few-shot in-context learning (Moslem et al., 2023). However, domain-specific or terminology-based knowledge distillation (Treviso et al., 2023) from a massive LLM to a compact task-oriented MT model can help boost efficiency at inference time while enhancing domain adaptation and terminology adherence. Obviously, when authentic in-domain data is available, it can be used for fine-tuning instead of synthetic data for domain adaptation of the MT model. In production workflows,

[6] https://github.com/huggingface/transformers

only segments that do not meet specific quality criteria are passed to either human or automatic post-editing. Hence, deployment of a model fine-tuned on in-domain data can reduce the number of translations that need post-editing.

2.3 Terminology-constrained Automatic Post-Editing

For the shared task, the organisers provided two term sets for each source sentence in the test and blind datasets, and expected the participants to generate two translations that use one term set each. In this step of terminology-constrained automatic post-editing, we aim to refine the translations generated by an MT system by inserting the required terminology. To this end, we checked the translations generated by the fine-tuned model from the previous step (cf. Section 2.2). For each term set provided for the sentence, if the translation does not include all the terms, we ran this step of terminology insertion into the translation.

This step involves instructing ChatGPT to post-edit the translation by making sure it includes all the terms without changing the rest of the translation. For the API's parameters, we used *top_p* 1 and *temperature* values 0 and 0.2, and then chose the generation that fixed more terms.

Example prompt: Terminology-constrained post-editing

In the following <tgt_lang> translation, use the <tgt_term> to translate the <src_lang> term <src_term>, and the...[7] Leave everything else the same.\n\n

<src_lang>: <src_segment>\n
<tgt_lang>: <tgt_segment>

3 Evaluation

To assess the effectiveness of our process, we conducted two types of evaluation: (i) term-level evaluation in order to measure the level of adherence to the required terminology, and (ii) sentence-level evaluation in order to see whether the process affected the quality of the overall translation.

3.1 Term-level Evaluation

In Tables 3 and 4, we report the number of terms used in the translations of the test and blind datasets, respectively, in respect to the two term sets provided by the organisers. The results show the effectiveness of our proposed process, increasing the integration of the required terms in the final translations of the blind dataset from an average of 36.67% with the baseline generic model to an average of 72.88% after the LLM-based post-editing, across the three language pairs. Interestingly, prompting an LLM to integrate the required terms into the translations generated by a fine-tuned encoder-decoder MT model was more effective than solely using the fine-tuned model.

Lang	System	Total [1]	Used [1]	Total [2]	Used [2]	Avg %
DE-EN	Baseline	432	291	317	168	60.18
	Fine-tuned	432	302	317	165	60.98
	Term APE	432	**397**	317	**239**	**83.65**
EN-CS	Baseline	550	221	313	139	42.30
	Fine-tuned	550	135	313	108	29.53
	Term APE	550	**466**	313	**283**	**87.57**
ZH-EN	Baseline	1779	498	1938	491	26.66
	Fine-tuned	1779	854	1938	570	38.71
	Term APE	1779	**1137**	1938	**886**	**54.81**
Avg. %	Baseline					43.05
	Fine-tuned					43.07
	Term APE					**75.34**

Table 3: For the test dataset, the number of terms used in the translations from the first term set [1] and the second term set [2]. According to the results, terminology-constrained automatic post-editing ("Term APE") using ChatGPT achieved the best adoption of the required terminology.

Lang	System	Total [1]	Used [1]	Total [2]	Used [2]	Avg %
DE-EN	Baseline	11357	4120	11202	4623	38.77
	fine-tuned	11357	4130	11202	4621	38.81
	Term APE	11357	**6257**	11202	**5893**	**53.85**
EN-CS	Baseline	10626	3964	10563	5122	42.90
	Fine-tuned	10626	3397	10563	4412	36.87
	Term APE	10626	**8727**	10563	**8681**	**82.16**
ZH-EN	Baseline	2892	1375	2908	265	28.33
	Fine-tuned	2892	1422	2908	970	41.26
	Term APE	2892	**2471**	2908	**2322**	**82.65**
Avg. %	Baseline					36.67
	Fine-tuned					38.98
	Term APE					**72.88**

Table 4: For the blind dataset, the number of terms used in the translations from the first term set [1] and the second term set [2]. According to the results, terminology-based automatic post-editing ("Term APE") using ChatGPT achieved the best adoption of the required terminology.

3.2 Sentence-level Evaluation

After the end of the submission phase, the organisers released the references for the participants to conduct automatic evaluation. The main purpose of this sentence-based evaluation process is to determine whether terminology integration affected the overall quality of translation. In general, as demonstrated in Table 4 and Table 5, this terminology-constrained automatic post-editing step significantly increased the inclusion of the necessary

[7] We can add more terms, if needed.

terms into the final translation while improving translation quality across the three language pairs.

For the automatic evaluation of each MT system, we used the BLEU (Papineni et al., 2002), chrF++ (Popović, 2017), and COMET (Rei et al., 2020) metrics. Since many of the Chinese-to-English segments in the blind dataset did not have two term sets, we evaluated only those that had two term sets (1629 segments out of 2640 segments). We observe that the evaluation scores of the Chinese-to-English translation task are much lower than those of the two other language pairs. This can be due to the literary nature of the blind dataset extracted from Chinese novels, which might be difficult for both the MT model and automatic evaluation metrics.

Lang	Count	System	BLEU	chrF++	COMET
DE-EN	2963	Baseline	19.81	48.04	21.81
		Fine-tuned	19.27	47.75	21.51
		Term APE [1]	32.36	60.84	40.25
		Term APE [2]	27.84	56.84	33.20
		Term APE Avg.	**30.10**	**58.84**	**36.73**
EN-CS	3005	Baseline	29.13	53.11	50.90
		Fine-tuned	24.54	49.14	33.78
		Term APE [1]	45.65	67.36	79.84
		Term APE [2]	37.88	61.19	63.64
		Term APE Avg.	**41.77**	**64.28**	**71.74**
ZH-EN	1629	Baseline	6.95	27.95	-50.90
		Fine-tuned	7.76	29.26	-38.83
		Term APE [1]	9.56	32.80	-18.96
		Term APE [2]	11.93	35.30	-13.51
		Term APE Avg.	**10.75**	**34.05**	**-16.24**

Table 5: Automatic evaluation of the overall translation quality across the three language pairs based on the blind dataset. The "Baseline" refers to the OPUS model without fine-tuning, while "Fine-tuned" refers to the model after domain adaptation with the bilingual terminology-based synthetic data generated by an LLM. Finally, the three last rows for each language pair refer to using ChatGPT for terminology-constrained automatic post-editing ("Term APE") of the MT output generated by the fine-tuned model. In other words, "Term APE [1]" indicates the results when the first term set was used to prompt ChatGPT to integrate terms of this set into the translation generated by the fine-tuned model, while "Term APE [2]" refers to using the second term set. Finally, "Term APE Avg." is the average of "Term APE [1]" and "Term APE [2]" for each language pair. Terminology-constrained automatic post-editing with ChatGPT achieves the best results across the three language pairs in terms of the overall translation quality. As reported in Table 4, the number of terms integrated after the automatic post-editing step also increased.

Moreover, it is worth noting that we used the English term while generating bilingual synthetic data (cf. Section 2.1) for the three language pairs. However, English is the target language for both Chinese-to-English and German-to-English language directions, while it is the source language for the English-to-Czech language direction. This can explain the performance degradation after the fine-tuning step in the English-to-Czech language direction (cf. Tables 4 and 5). In other words, it is recommended in the step of bilingual synthetic data generation to either use the target term or both the source and target terms while prompting the LLM to generate translation pairs.

As explained in Section 2.3, our final step of terminology-constrained automatic post-editing involves instructing an LLM to insert terms that were missing from the output of the fine-tuned model. This significantly increased term usage across all the Chinese-to-English, English-to-Czech, and German-to-English language pairs (cf. Table 4). Furthermore, as demonstrated in Table 5, this step had no detrimental effects on translation quality. In fact, integrating the necessary terms into the translation using ChatGPT improved translation quality according to our automatic evaluation.

4 Conclusion and Future Work

In this work, we showed that applying a multistep process of mixed fine-tuning on terminology-based synthetic bilingual data and then terminology-constrained automatic post-editing with an LLM can increase the adherence to the pre-approved terms in the generated translations. By the end of the process, the use of the required terms has increased in the translations of the blind dataset across the three language pairs from an average of 36.67% with the baseline generic model to an average of 72.88% after instructing an LLM to integrate the required terms into the translations.

Due to the task restrictions, we had to fine-tune OPUS models only. We would like to experiment with fine-tuning NLLB models, and probably the new SeamlessM4T (Barrault et al., 2023), Mistral (Jiang et al., 2023), and MADLAD-400 models (Kudugunta et al., 2023), on the same data and compare the output quality. In our experiments, we employed ChatGPT "gpt-3.5-turbo" for both terminology-based synthetic data generation and terminology-constrained automatic post-editing, as it is a relatively efficient and cost-effective option. In the future, we would like to repeat the same experiments with GPT-4 in order to assess the benefit of using a stronger language model on overall performance. We observe that BLOOM can be used as an alternative LLM for data generation; however, one-shot generation might work better than zero-shot generation. In this case, the prompt can consist of a term, a bilingual sentence pair,

and then another term. Interestingly, the model will predict a new translation pair including the second term. While certain open-source models such as Llama 2 and Falcon might be employed for the terminology-constrained automatic post-editing step for certain languages, we suspect that they will need fine-tuning before being reliably usable for most languages.

In future work, we will carry out a deeper analysis of the generated synthetic data together with the outputs of the fine-tuned models in order to understand how the properties of the synthetic data affect the fine-tuning results. It is important also to test the same approach for other languages, especially low-resource language pairs.

Moreover, it would be interesting to exclude the fine-tuning step and assess the overall translation quality after LLM-based post-editing. It is possible that domain adaptation through fine-tuning the baseline MT model either on authentic or synthetic data would still be beneficial. It can lead to domain-specific improvements in the overall translation quality that may not be achievable by the baseline model or the terminology-constrained post-editing step. Again, deploying a model fine-tuned on in-domain data into production can enhance terminology adherence in initial translations. As there is no need to send the translations that already include the pre-approved terms to the LLM for terminology-constrained post-editing, this can reduce the number of translations that require post-editing. Such an efficient workflow can allow us to save resources, and minimise latency at inference time. Similarly, there are potential advantages of employing an LLM for post-editing rather than for direct translation. Instead of solely relying on the translation quality of the LLM, quality estimation can be performed to select the best MT model in general or for the current source text segment. Ultimately, only segments that do not meet quality criteria are then passed to the LLM for post-editing.

Acknowledgements

This work is supported by the Science Foundation Ireland (SFI) Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224, the ADAPT Centre for Digital Content Technology under SFI's Grant No. 13/RC/2106_P2, and Microsoft Research.

References

Loic Barrault, Andy Chung, David Dale, Ning Dong (ai), Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Peng-Jen Chen, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Abinesh Ramakrishnan, Alexandre Mourachko, Amanda Kallet, Ann Lee, Anna Sun, Bapi Akula, Benjamin Peloquin, Bernie Huang, Bokai Yu, Brian Ellis, Can Balioglu, Carleigh Wood, Changhan Wang, Christophe Ropers, Cynthia Gao, Daniel Li (fair), Elahe Kalbassi, Ethan Ye, Gabriel Mejia Gonzalez, Hirofumi Inaguma, Holger Schwenk, Igor Tufanov, Ilia Kulikov, Janice Lam, Jeff Wang (pm Ai), Juan Pino, Justin Haaheim, Justine Kao, Prangthip Hasanti, Kevin Tran, Maha Elbayad, Marta R Costa-jussa, Mohamed Ramadan, Naji El Hachem, Onur Çelebi, Paco Guzmán, Paden Tomasello, Pengwei Li, Pierre Andrews, Ruslan Mavlyutov, Russ Howes, Safiyyah Saleem, Skyler Wang, Somya Jain, Sravya Popuri, Tuan Tran, Vish Vogeti, Xutai Ma, and Yilin Yang. 2023. SeamlessM4T—Massively Multilingual & Multimodal Machine Translation.

Rachel Bawden and François Yvon. 2023. Investigating the Translation Performance of a Large Multilingual Language Model: the Case of BLOOM. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170, Tampere, Finland. European Association for Machine Translation.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*, volume 33, pages 1877–1901, Virtual. Curran Associates, Inc.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M Dai, Thanumalayan Sankaranarayana

Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. *arXiv preprint arXiv:2204.02311 [cs.CL]*.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2017. An Empirical Comparison of Domain Adaptation Methods for Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–391, Vancouver, Canada. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672 [cs.CL]*.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training Neural Machine Translation to Apply Terminology Constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

Miriam Exel, Bianka Buschbeck, Lauritz Brandt, and Simona Doneva. 2020. Terminology-Constrained Neural Machine Translation at SAP. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 271–280, Lisboa, Portugal. European Association for Machine Translation.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based Phrase-level Prompting of Large Language Models for Machine Translation. *arXiv preprint arXiv:2302.07856 [cs.CL]*.

Rejwanul Haque, Yasmin Moslem, and Andy Way. 2020. Terminology-Aware Sentence Mining for NMT Domain Adaptation: ADAPT's Submission to the Adap-MT 2020 English-to-Hindi AI Translation Shared Task. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON): Adap-MT 2020 Shared Task*, pages 17–23, Patna, India. NLP Association of India (NLPAI).

Eva Hasler, Adrià de Gispert, Gonzalo Iglesias, and Bill Byrne. 2018. Neural Machine Translation Decoding with Terminology Constraints. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 506–512, New Orleans, Louisiana. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. *arXiv preprint arXiv:2302.09210 [cs.CL]*.

Chris Hokamp and Qun Liu. 2017. Lexically Constrained Decoding for Sequence Generation Using Grid Beam Search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

J Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme. 2019a. Improved Lexically Constrained Decoding for Translation and Monolingual Rewriting. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 839–850, Minneapolis, Minnesota. Association for Computational Linguistics.

Junjie Hu, Mengzhou Xia, Graham Neubig, and Jaime Carbonell. 2019b. Domain Adaptation of Neural Machine Translation by Lexicon Induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2989–3001, Florence, Italy. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825 [cs.CL]*.

Wenxiang Jiao, Wenxuan Wang, Jen-Tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine. *arXiv preprint arXiv:2301.08745 [cs.CL]*.

Marcin Junczys-Dowmunt. 2018. Dual Conditional Cross-Entropy Filtering of Noisy Parallel Corpora. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 888–895, Belgium, Brussels. Association for Computational Linguistics.

Guillaume Klein, Dakun Zhang, Clément Chouteau, Josep Crego, and Jean Senellart. 2020. Efficient and

high-quality neural machine translation with OpenNMT. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 211–217, Stroudsburg, PA, USA. Association for Computational Linguistics.

Catherine Kobus, Josep Crego, and Jean Senellart. 2017. Domain Control for Neural Machine Translation. In *Proceedings of Recent Advances in Natural Language Processing*, pages 372–378, Varna, Bulgaria.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A Multilingual And Document-Level Large Audited Dataset. *arXiv preprint arXiv:2309.04662 [cs.CL]*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J Mielke, Wilson Y Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán,

Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv preprint arXiv:2211.05100 [cs.CL]*.

Elise Michon, Josep Crego, and Jean Senellart. 2020. Integrating Domain Terminology into Neural Machine Translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3925–3937, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Robert C Moore and William Lewis. 2010. Intelligent Selection of Language Model Training Data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden. Association for Computational Linguistics.

Yasmin Moslem, Rejwanul Haque, John Kelleher, and Andy Way. 2022. Domain-Specific Text Generation for Machine Translation. In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 14–30, Orlando, USA. Association for Machine Translation in the Americas.

Yasmin Moslem, Rejwanul Haque, John D Kelleher, and Andy Way. 2023. Adaptive Machine Translation with Large Language Models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

Prashanth Nayak, Rejwanul Haque, John D Kelleher, and Andy Way. 2023. Instance-Based Domain Adaptation for Improving Terminology Translation. In *Proceedings of Machine Translation Summit XIX: Research Track*, pages 222–231, Macau SAR, China. Association for Machine Translation in the Americas.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS 2022)*, pages 27730–27744, New Orleans, Louisiana, USA. Curran Associates, Inc.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only. *arXiv preprint arXiv:2306.01116 [cs.CL]*.

Alberto Poncelas, Gideon Maillette de Buy Wenniger, and Andy Way. 2019. Adaptation of Machine Translation Models with Back-Translated Data Using Transductive Data Selection Methods. In *Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing CICLing 2019: Computational Linguistics and Intelligent Text Processing*, pages 567–579, La Rochelle, France. Springer Nature Switzerland.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post and David Vilar. 2018. Fast Lexically Constrained Decoding with Dynamic Beam Allocation for Neural Machine Translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Alham Fikri Aji, Zhengzhong Liu, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Preslav Nakov, Timothy Baldwin,

and Eric Xing. 2023. Jais and Jais-chat: Arabic-Centric Foundation and Instruction-Tuned Open Generative Large Language Models. *arXiv preprint arXiv:2308.16149 [cs.CL]*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mGPT: Few-Shot Learners Go Multilingual. *arXiv preprint arXiv:2204.07580 [cs.CL]*.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288 [cs.CL]*.

Marcos Treviso, Ji-Ung Lee, Tianchu Ji, Betty van Aken, Qingqing Cao, Manuel R Ciosici, Michael Hassid, Kenneth Heafield, Sara Hooker, Colin Raffel, Pedro H Martins, André F T Martins, Jessica Zosa Forde, Peter Milder, Edwin Simpson, Noam Slonim, Jesse Dodge, Emma Strubell, Niranjan Balasubramanian, Leon Derczynski, Iryna Gurevych, and Roy Schwartz. 2023. Efficient methods for natural language processing: A survey.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems (NIPS 2017)*, volume 30. Curran Associates, Inc.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for Translation: Assessing Strategies and Performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. Github (mesh-transformer-jax).

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual Universal Sentence Encoder for Semantic Retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.

OPUS-CAT Terminology Systems for the WMT23 Terminology Shared Task

Tommi Nieminen
University of Helsinki
tommi.nieminen@helsinki.fi

Abstract

This paper describes the submission of the OPUS-CAT project to the WMT 2023 terminology shared task. We trained systems for all three language pairs included in the task. All systems were trained using the same training pipeline with identical methods. Support for terminology was implemented by using the currently popular method of annotating source language terms in the training data with the corresponding target language terms.

1 Introduction

OPUS-CAT (Nieminen, 2021) is a collection of open source software consisting of a local neural machine translation (NMT) engine and plugins for computer-assisted translation (CAT) tools, such as Trados, memoQ and OmegaT. OPUS-CAT enables the use of NMT models trained in the OPUS-MT project (Tiedemann and Thottingal, 2020) in professional translation. As OPUS-CAT is aimed at professional translators, it is designed to be integrated into normal translation workflows. Multilingual term bases are one part of those workflows, so we have decided to implement a functionality for utilizing term bases in OPUS-CAT. This paper describes the methods used in OPUS-CAT for enforcing the use of terminology in machine translation output and the results of applying these methods to the data provided in the shared task. We trained new models for all three language pairs in the shared task. The shared task results were not available at the time of the submission of this paper.

2 Related work

Most published methods of constraining an NMT model to generate terminologically correct translations fall into three categories.

2.1 Constrained decoding

Hokamp and Liu (2017); Hasler et al. (2018): The beam search algorithm is modified to enforce the generation of target terms for each source term identified in the source sentence. The main advantage of constrained decoding is that it can be used with any model. The main disadvantages are slower decoding speed, and quality degradation due to the unconditional prioritizing of target terms, even in inappropriate contexts (such as generating the target term multiple times in the translation).

2.2 Pass-through term placeholders

Michon et al. (2020): Source terms identified in the source sentence are replaced by placeholders, which the NMT model passes through to the translation. The placeholders generated in the translation are then replaced by corresponding target terms. In order for the model to learn the correct pass-through behaviour, the model has to be trained with data that has been augmented with sentence pairs containing aligned placeholders on source and target sides. The main advantage of this approach is that the target terms are usually generated in correct positions. The disadvantage is that the information in the source term is discarded, which may degrade the quality of the overall translation. Generating morphological features for the target term may also be difficult.

2.3 Injecting target terms as soft constraints

Dinu et al. (2019): Source terms identified in the source sentence are annotated with target term information, and the NMT model uses these target term annotations to generate the term translations. Similar to the pass-through placeholder method, the training data of the model needs to be augmented with sentence pairs, where the source sentence has been annotated with target term information that also occurs in the target sentence. This will induce the model to generate translations that conform to the target term information present in the source text. While the constrained decoding and pass-through placeholder methods unconditionally

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 912–918
December 6–7, 2023.

ally enforce the use of the specified terminology in the generated translation (they place hard constraints on the output), in this method terms are soft constraints on the output: contextual factors may cause the model to not use the specified term in the translation. This is the desired behaviour, since terms are often polysemous, and the specified term translation is usually only appropriate for one sense of the term. For instance, a terminology might specify a translation for the word *file*, but the translation would only be relevant for the sense of *file* meaning an individual file in a computer file system, instead of e.g. a physical file, a wood file, or the imperative of the verb *to file*.

The terminology support in OPUS-CAT is based on the soft constraint method as it is the simplest to implement and has performed best in previous evaluations (Alam et al., 2021b).

3 Model training

The models were trained using a modified version of Mozilla's *firefox-translations-training*[1], an end-to-end pipeline for building NMT models, based on the Snakemake workflow management system (Mölder et al., 2021). The pipeline loads, preprocesses, cleans and filters the training data, and trains and evaluates the NMT models. For this shared task, a terminology annotation workflow has been added to the pipeline[2].

3.1 Data

The models were trained using the data provided for the constrained track of WMT23. Since sufficient parallel data was available for each language pair, we did not include any back-translated monolingual data in the training corpus. This simplifies and speeds up training, and from the point of view of terminological correctness there does not appear to be any obvious benefit to using back-translated data, even though it would almost certainly increase general output quality.

3.2 Data cleaning

The data was cleaned and filtered using the standard *firefox-translations-training* workflow, which consists of monolingual cleaning of source and target corpora, followed by the filtering of parallel sentences with Bicleaner or Bicleaner-AI. Data for **en-cs** and **de-en** were filtered with Bicleaner-AI, while no parallel cleaning was performed for **zh-en**, as no Bicleaner-AI model for **zh-en** was available to the pipeline.

3.3 Terminology annotation

A part of the cleaned and filtered data was annotated with artificial term information (the annotation script is available from https://github.com/TommiNieminen/soft-term-constraints). First, artificial term data is generated from the parallel data:

1. **POS tagging and dependency parsing:** Stanza (Qi et al., 2020) was used to identify the parts-of-speech (POS) and dependency relations of the tokens in the source and target sentences.

2. **Chunking:** The POS and dependency data from step 1 was used to identify noun and verb phrase chunks in the source and target sentences.

3. **Word alignment:** The filtered parallel corpus was aligned on word-level using FastAlign (Dyer et al., 2013).

4. **Chunk alignment:** Source chunks that were aligned with target chunks were identified based on the word alignment from step 3.

The above method is identical to the one in Bergmanis and Pinnis (2021) except for the addition of chunking.

As analyzing sentences with Stanza is quite slow, only a small portion of the parallel data was analyzed (approximately one in ten sentences). The noun and verb phrase chunks identified on the basis of the analysis were saved and used to annotate the data using two different annotation methods (see table 2 for examples):

1. **Append**: The target language chunk was appended to the aligned source language chunk, with the source and target chunks separated with a special separator tag. A start tag was also added before the start of the source chunk, and an end tag was added after the end of the target chunk.

2. **Replace**: The source language chunk was replaced with the aligned target language chunk. The target chunk in the source sentence was tagged with start and end tags.

[1] https://github.com/mozilla/firefox-translations-training

[2] https://github.com/GreenNLP/firefox-translations-training/tree/develop

Language pair	Raw	Cleaned	Annotated
Chinese to English	35,452,884	28,840,867	2,884,058
German to English	294,331,299	182,977,635	18,297,581
English to Czech	56,288,239	35,046,151	2,704,588

Table 1: Amount of parallel sentences available for each language pair. Base model is trained with cleaned data, and the terminology models are fine-tuned with a combination of clean and annotated data or just annotated data (**-omit** models).

Source	`This product is no longer available`
Append	`This <term_start> product <term_end> produkt <trans_end> is no longer available.`
Replace	`This <term_start> produkt <term_end> is no longer available.`

Table 2: Examples of **append** and **replace** annotation methods

These methods are identical to the ones in Dinu et al. (2019) except for the use of tags instead of factors to identify terms (similar to Ailem et al. (2021).

Since a source sentence can potentially have any number of source terms, the training data needs to contain source sentences with different amounts of annotated terms. The annotation algorithm keeps track of how many sentences with n terms have been annotated so far, and tries to ensure that the sentence counts approximate a geometric series, where the amount of sentences gets halved for every extra term. For instance, the annotated corpus for en-cs contains 1,353,810 sentences with one term, 676,895 sentences with two terms, 338,414 sentences with three terms and so forth. The justification for the ratio is that most sentences will contain only few terms, so the lower counts should be emphasized in training.

4 Observations on the shared task

This year's terminology task differs in from real-world use of terminology in machine translation in two important aspects:

1. Source terms have been unambiguously identified.

2. Target terms are specified in an already inflected form. This inflected form has been extracted from a reference translation, and therefore has a high probability of being a correct form to use in a translation.

In actual use cases, the NMT system would have to identify the source terms based on a lemma form provided in a term base, and only the lemma form of the target term would be available. The probability of the lemma term occurring as such in a correct translation is much lower than for the inflected term from a reference translation. The shared task is therefore much easier than the real-world task of translating with a term base.

Due to the use of inflected terms, the shared task also favours soft constraint models where the model is trained on surface forms of terms instead of lemma forms. Because of this, the models we have submitted for the shared task all use surface forms of the terms. However, this will induce the models to learn a simple copy behaviour (Dinu et al., 2019), instead of the more desirable copy-and-inflect behaviour (Bergmanis and Pinnis, 2021). In our OPUS-CAT production models, we intend to use lemma-based constraints, since we expect them to perform better in real-world scenarios, especially with morphologically complex target languages.

5 Models

Five different models were trained for each language pair. All of the models were trained with Marian (Junczys-Dowmunt et al., 2018) using the **transformer-big** model architecture (Vaswani et al., 2017). For each language pair, a combined SentencePiece (Kudo and Richardson, 2018) vocabulary (32,000 symbols, out of which ten symbols were reserved as potential term tags by using the user-defined symbol functionality of SentencePiece) was trained and used for both source and target languages. As transformer-big models are costly to train, a single base model was trained for each language pair using just the filtered corpus, and the base model was then fine-tuned with data

that had been augmented with the terminological annotations. Another motivation for using fine-tuning is the reuse of models: OPUS-CAT uses the OPUS-MT model collection that contains thousands of pre-trained models, and fine-tuning those models to support the use of terminology instead of training terminology models from scratch saves time and resources.

Yet another advantage of fine-tuning is that it makes it possible to quickly test the performance of different term annotation schemes. As mentioned, we experimented with the **append** and **replace** methods. For both methods, two models were trained, one where the annotated sentences were combined with the unannotated sentences when fine-tuning (**add**), and one where the unannotated sentences were omitted (**omit**). The expectation is that the **omit** model will specialize better to term translation, while the **add** model will retain better generic translation capabilities. In production use it may be best to use a specialized term model when terms are detected in the source sentence, and revert back to a generic model when no terms are detected.

The **zh-en** base model was trained until convergence (chrF validation metric did not improve for 20 consecutive validation steps). For the **en-cs** and **de-en** base models the training did not have time to converge before the deadline for shared task submission, but both models were trained sufficiently long to obtain competitive evaluation scores (on par with scores published for existing OPUS-MT models). The terminology models were trained by fine-tuning the base model with annotated data for one epoch.

When translating with a terminology model and a term base, a script is used to identify terms in the source text and to annotate the terms in the source sentence before translation, using the same annotation scheme as in the training data. Since the target side of the training data was not modified, the translation does not need to be post-processed.

5.1 Model n-best combination and reranking

For the submission to the shared task, we combine the outputs of the different types of models using a simple n-best reranking method (this is referred to as the **mixture** model in the tables):

1. An n-best list of size 8 is generated for each source sentence by each model.
2. Term occurrences are counted for each translation in the n-best lists.
3. The translation containing the most terms in all n-best lists is chosen as the final translation.
4. If translations from different models have the same amount of terms, the final translation is picked based on the following model hierarchy: **base**, **append**, **replace**, **append-omit**, **replace-omit** (the assumption is that the quality is best for the base model and worst for the omit models).
5. If there are multiple translations with the same amount of terms in a model's n-best list, translations higher in the n-best list are preferred.

The motivation for using this reranking method is that since the models use different approaches to generate translations, their combined n-best lists will be diverse, which increases the probability of finding a translation with correct terms. Also, in general it makes sense to rerank n-best lists in terminology translation, since the criteria for reranking is so clear (the highest amount of term occurrences).

6 Evaluation

6.1 Evaluation methods

General model performance was evaluated with BLEU and chrF metrics using sacreBLEU (Post, 2018).

Terminological correctness was evaluated by simply counting what percentage of the specified terms actually occur in the translation in the surface form in which they are defined. This naive method ignores two important issues: the correct placement of the term within the translation, and the matching of all other inflected forms of the term. Alam et al. (2021a) introduces more sophisticated term accuracy metrics to alleviate these issues, but we decided against applying them. Since we use evaluation mainly for sanity checking soft constraint models, which generally place terms correctly (and do not place terms at all if no plausible position is found for them), evaluating the correct placement is not crucial. Likewise, matching all inflected forms is not crucial in the context of this shared task, since the terminology is provided in an already inflected form, and our models have been trained with surface term annotations, and will likely have learned to copy the single inflected form provided to them.

DE-EN	flores-dev	wmt13	wmt16	wmt18	wmt20
base	**37.6/64.7**	**32.4/59.1**	**34.7/61.0**	32.5/58.6	23.3/51.6
append	**37.6**/64.6	**32.4**/58.9	34.5/60.9	32.5/58.5	23.0/51.6
append-omit	37.1/64.5	32.3/59.0	34.5/60.9	**32.6/58.7**	21.8/49.9
replace	**37.6**/64.6	32.1/58.7	34.3/60.7	32.3/58.4	**23.5/51.8**
replace-omit	37.2/64.5	32.2/58.9	34.2/60.7	32.5/58.6	22.0/50.4

ZH-EN	flores	wmt20	wmt21	wmt22
base	25.9/55.8	25.7/55.7	20.4/50.2	18.6/48.6
append	**27.2/56.7**	**27.8/57.3**	**22.2**/51.7	**19.9/49.9**
append-omit	26.8/56.2	27.0/56.5	21.6/51.0	19.3/49.2
replace	27.0/56.6	27.7/57.1	**22.2/51.8**	19.5/49.6
replace-omit	26.9/56.3	27.0/56.6	21.6/51.1	19.2/49.1

EN-CS	flores	wmt13	wmt16	wmt18	wmt20
base	**34.1/60.6**	**27.0/53.5**	**29.3/56.6**	**24.2/52.3**	20.5/**50.4**
append	33.4/60.1	26.8/53.3	29.2/56.5	23.8/51.9	**20.6/50.4**
append-omit	33.6/60.3	**27.0**/53.3	29.0/56.3	24.0/52.0	19.7/49.4
replace	33.5/60.3	26.8/53.4	29.2/56.5	24.1/52.1	20.4/50.2
replace-omit	33.6/60.2	26.8/53.2	29.0/56.2	23.9/51.9	20.2/49.7

Table 3: General translation performance measured as BLEU/chrF. Note that the input to the term models was not annotated with terms when translating these test sets, they translated the same unannotated input as the base model. Therefore it is to be expected that the term models perform worse in this evaluation.

		Exact term accuracy
DE-EN (100)	base	0.618
	append	0.911
	append-omit	0.854
	replace	0.886
	replace-omit	0.902
ZH-EN (100)	base	0.367
	append	0.933
	append-omit	0.933
	replace	0.900
	replace-omit	0.967
EN-CS (100)	base	0.496
	append	0.837
	append-omit	0.756
	replace	0.829
	replace-omit	0.772

Table 4: Term translation accuracy with the shared task dev set (sentence count is in parentheses under the language pair). In this scenario, the terms have been annotated to the input of the term models, and the term models perform better than the base model, as is to be expected.

6.2 Evaluation data

Models were evaluated against a selection of test sets allowed for the constrained track of WMT23 (see table 3 for results). Terminological correctness was evaluated using the development sets provided in the shared task (see table 4 for results). As the shared task development sets were quite small, we also created artificial terminology test sets for each language pair from the constrained track test sets, using the same annotation script that was used to annotate the training data (we did not use pre-existing terminologically annotated corpora due to the constrained track restrictions). Aligned noun and verb phrase chunks were identified in the test set sentences, and converted into sentence-level dictionaries similar to those in the shared task development sets (see table 5 for results).

Most NMT models trained on parallel data will exhibit some degree of copy behaviour, since source texts often contain target language words (this is especially common when the target language is English, due to its dominant position as a world language). Therefore it is plausible that the base models are already capable of copying target terms injected into the source sentence to the

DE-EN (6550)	Exact term accuracy	BLEU/ chrF
base	0.732	42.5/65.9
append	0.973	46.8/**69.2**
append-omit	0.942	**46.9/69.2**
replace	**0.977**	46.7/**69.2**
replace-omit	0.958	46.8/**69.2**
mixture	*0.997*	46.4/69.1
base-term	0.945	44.3/67.9

ZH-EN (5687)	Exact term accuracy	BLEU/ chrF
base	0.656	22.9/53.1
append	**0.949**	**26.1**/55.9
append-omit	0.899	24.9/54.8
replace	0.940	25.9/55.9
replace-omit	0.892	25.0/55.0
mixture	*0.985*	**26.1/56.2**
base-term	0.884	22.7/53.6

EN-CS (8204)	Exact term accuracy	BLEU/ chrF
base	0.651	28.3/55.5
append	0.902	31.4/58.4
append-omit	0.803	30.2/57.3
replace	**0.909**	31.2/58.3
replace-omit	0.861	30.2/57.6
mixture	*0.959*	**32.0/59.0**
base-term	0.827	29.0/56.8

Table 5: Term translation accuracy with the artificial term test set (test set sentence count is in parentheses under the language pair). Note that **mixture** will always have the best term accuracy, since it combines the output of other models based on term accuracy. Target terms have been added to the input for all models expect **base**. **base-term** is a **base** model translating input where source terms have been replaced with target terms.

translation. To determine the extent of this innate copying ability of the base model and the actual improvement brought by fine-tuning, a separate **base-term** test set was created from the artificial term test set by replacing the source terms in the source sentences with corresponding target terms.

6.3 Interpretation of the evaluation results

Results of the evaluation mostly conform to expectations. All soft constraint models outperform the base model in term translation, with the **append** and **replace** models performing best. This is somewhat surprising, since the **append-omit** and **replace-omit** models were expected to specialize better to term translation.

It is also surprising that the general translation quality of the soft constraint models is comparable to that of the base models. Strangely, the **zh-en** soft constraint models clearly outperform the base model even in general translation. This may be due to the **zh-en** base model converging early, after only 6 epochs of training. Still, it is counter-intuitive that fine-tuning with the small **omit** data sets consisting only of annotated sentences should noticeably improve general translation quality.

The results also confirm that the base models are quite capable of copying exact terms from the input sentence into the translation, especially the **de-en** model. However, injecting terms directly into the base model input seems to noticeably lower the overall translation quality.

7 Conclusion

Our submission for the shared task confirms that soft terminology constraint methods work with a variety of language pairs. We also demonstrate that soft constraint models can be created by fine-tuning base transformer models, which speeds up training and the investigation of different soft constraint methods and parameters. The results also indicate that fine-tuned soft constraint models have acceptable general translation quality, and do not require a back-off base model in production use.

Limitations

The soft constraint methods discussed assume terms are inflected, which is not usually the case when actually working with term bases. This limits the usability of the methods, especially with morphologically complex target languages. However, the annotation script also supports the use of lemma forms of terms. The reranking method used to produce the best term accuracy is computationally heavy, as it requires decoding with five separate **transformer-big** models.

Acknowledgements

The work reported in this paper has been funded and supported by the Swedish Culture Foundation in Finland and by the Finnish Academy. Computational resources were provided by CSC – IT Center for Science, Finland.

References

Melissa Ailem, Jingshu Liu, and Raheel Qader. 2021. Encouraging neural machine translation to satisfy terminology constraints. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1450–1455, Online. Association for Computational Linguistics.

Md Mahfuz Ibn Alam, Antonios Anastasopoulos, Laurent Besacier, James Cross, Matthias Gallé, Philipp Koehn, and Vassilina Nikoulina. 2021a. On the evaluation of machine translation for terminology consistency. *CoRR*, abs/2106.11891.

Md Mahfuz Ibn Alam, Ivana Kvapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Kweonwoo Jung, Philipp Koehn, and Vassilina Nikoulina. 2021b. Findings of the wmt shared task on machine translation using terminologies. In *Proceedings of the Sixth Conference on Machine Translation*, pages 652–663, Online. Association for Computational Linguistics.

Toms Bergmanis and Mārcis Pinnis. 2021. Facilitating terminology translation with target lemma annotations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3105–3111, Online. Association for Computational Linguistics.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 644–648. The Association for Computational Linguistics.

Eva Hasler, A. Gispert, Gonzalo Iglesias, and Bill Byrne. 2018. Neural machine translation decoding with terminology constraints. In *North American Chapter of the Association for Computational Linguistics*.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Annual Meeting of the Association for Computational Linguistics*.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Elise Michon, Josep Maria Crego, and Jean Senellart. 2020. Integrating domain terminology into neural machine translation. In *International Conference on Computational Linguistics*.

Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B. Hall, Christopher H. Tomkins-Tinch, Vanessa V. Sochat, Jan Forster, Soohyun Lee, Sven O. Twardziok, Alexander Kanitz, Andreas Wilm, Manuel Holtgrewe, Sven Rahmann, Sven Nahnsen, and Johannes Köster. 2021. Sustainable data analysis with snakemake. *F1000Research*, 10:33.

Tommi Nieminen. 2021. OPUS-CAT: Desktop NMT with CAT integration and local fine-tuning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 288–294, Online. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

VARCO-MT: NCSOFT's WMT'23 Terminology Shared Task Submission

Geon Woo Park, Junghwa Lee, Meiying Ren, Allison Shindell, Yeonsoo Lee
NCSOFT NLP Center
{parkku01, jleehhh0217, mia1211, shindell, yeonsoo}@ncsoft.com

Abstract

A lack of consistency in terminology translation undermines quality of translation from even the best performing neural machine translation (NMT) models, especially in narrow domains like literature, medicine, and video game jargon. Dictionaries containing terminologies and their translations are often used to improve consistency but are difficult to construct and incorporate. We accompany our submissions to the WMT '23 Terminology Shared Task with a description of our experimental setup and procedure where we propose a framework of terminology-aware machine translation. Our framework comprises of an automatic terminology extraction process that constructs machine translation datasets with terminology dictionaries in low-supervision settings and two model architectures with terminology constraints. Our models outperform baseline models by 21.51%p and 19.36%p in terminology recall respectively on the Chinese to English WMT'23 Terminology Shared Task test data.

1 Introduction

The WMT'23 Terminology Shared Task aims to assess machine translation models' abilities to leverage additional information. A terminology dictionary is provided with each line of source text. This is particularly useful for terminology consistency. The WMT'23 Terminology Shared task includes the language pairs Chinese-English, English-Czech, and German-English. We focus our submission on the Chinese-English pair. The task consists of three modes, as shown in Table 1.

Mode 1 assesses translation quality without additional terminology information. Mode 2 assesses translation quality with additional terminology information. Mode 3 assesses translation quality with a glossary containing random non-terminology.

In this paper, we describe our model building process for terminology translation from data preprocessing to model evaluation. We present two Transformer-based encoder-decoder models: Terminology Self-selection Neural Machine Translation (TSSNMT) and ForceGen Transformer (ForceGen-T). TSSNMT uses a shared encoder with a gating mechanism (Bapna and Firat, 2019), allowing the model to determine the weights of the source sentence and terminologies to use during generation. ForceGen-T enforces a decoder to generate the terminologies via force decoding (Reheman et al., 2023) and copy mechanism (Song et al., 2019), which enables the model to attend to terminologies during generation. Both models significantly outperform the baseline model.

2 Related works

Previous work on enhancing machine translation with pre-defined terminology encompasses three primary approaches.

First, a data-driven approach where terminologies are appended to input sentences (Dinu et al., 2019; Song et al., 2019). Song et al. (2019) suggest using copy mechanism to instruct the model to replicate the target terminology during the generation process.

Second, an alternative approach focuses on manipulating the model architecture. Bapna and Firat (2019) have used input sentences and their corresponding retrieved translation pairs to encode conditional source target memory. This approach uses a gated multi-source attention mechanism, which takes the encoded representation and the hidden state of the source as input, thereby steering the model toward the generation of the intended translation.

Third, efforts have been directed at tailoring the decoding process to incorporate terminologies. Hokamp and Liu (2017) and Post and Vilar (2018) have introduced constrained decoding techniques that reinforce the translated output's pre-specified terminologies.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 919–925

Mode	Source Input	Glossary Input	Target Output
1	脱离这些影响的建筑, 17世纪的建筑师伊尼戈琼斯和 克里斯托弗·雷恩牢固确立了 在英国的古典主义.	-	Architects Inigo Jones and Christopher Ren strongly established classicalism in England in the 17th century, free from these influences.
2		{"en": "Christopher Wren", "zh": "克里斯托弗·雷恩"}	Inigo Jones and **Christopher Wren**, two architects from the 17th century, strongly established classical architecture in England, free from these influences.
3		["en": "firmly", "zh": "牢固", "en": "free", "zh": "脱离"]	Building designs that were **free** from these influences, 17th century architects Inigo Jones and Christopher Ren **firmly** established classical architecture in England.

Table 1: Different Mode Scenarios

3 Data Process

The WMT'23 Terminology Shared Task is a constrained track, following the same rules of data usage as the WMT'23 General MT Task, forbidding the use of external data. However, unlike the WMT'21 Terminology Task, the provided training data lacks terminology information, and need to be artificially constructed.

3.1 Data Filtering

We first filter noisy data. We referenced the data cleaning methods described in the WMT'21 Terminology submissions (Molchanov et al., 2021; Wang et al., 2021).

1. Remove pairs that contain sentences that:
   (a) are empty, too short, too long.
   (b) are contain only symbols.
   (c) are at least 3 times longer than their counterpart.
2. Delete text pairs identified to be the wrong language. We used a combination of our in-house language detector for short texts and LangID (Lui and Baldwin, 2012) for long texts.
3. Remove pairs outside of a selected cosine similarity scope of latent vectors constructed by the LaBSE model (Feng et al., 2022).

See Appendix A for the amount of data filtered and the resulting performance comparison.

3.2 Word Alignment

After filtering data, we tokenize and word-align the text to extract desired terminology pairs for Modes 2 and 3. The overall process is described in Figure 1. We use our in-house tokenizer, referring to the tagging schema from Luo et al. (2019) for Chinese and the Moses (Koehn et al., 2007) tokenizer for English. Next, the tokenized parallel data is fed into a LaBSE (Feng et al., 2022) based word aligner, AccAlign (Wang et al., 2022). AccAlign generates pairs of indices for words from the source and target text, which are then utilized in extracting terminology pairs.

3.3 Terminology Extraction: Mode 2

The terminology extracted for Mode 2 are named entities, excluding time and number expressions. We use `SpaCy`'s (Honnibal and Montani, 2017) `zh_core_web_lg` as the Chinese NER model and `en_core_web_md` as the English NER model. Furthermore, we consider Chinese four-character idioms extracted by our in-house tokenizer. The idioms are added as additional Mode 2 candidates. Next, we reference our word alignment results from Section 3.2 to map candidates to their corresponding targets.

AccAlign occasionally fails to align multi-word terminology completely, which poses an issue for Chinese idioms. To account for this, we implement a soft matching strategy to interpret AccAlign's output indices, where we extract the entire phrase if the aligner maps the beginning and end indices, even when alignment is not complete in the middle of the phrase. We use strict matching for named entities, which only extracts words that appear in the alignment results.

To guarantee that our extracted terminology is accurate and exhaustive, we repurpose the provided training data source WikiTitles as an additional resource for terminology pairs. For each pair in WikiTitles, we check whether a term and its translation were present on both sides of the parallel text and add the relevant term pairs into the Mode 2 glossary.

3.4 Terminology Extraction: Mode 3

The terminology extracted for Mode 3 is intended to be relatively random yet accurate pairs from the parallel text. For simplicity, we exclusively

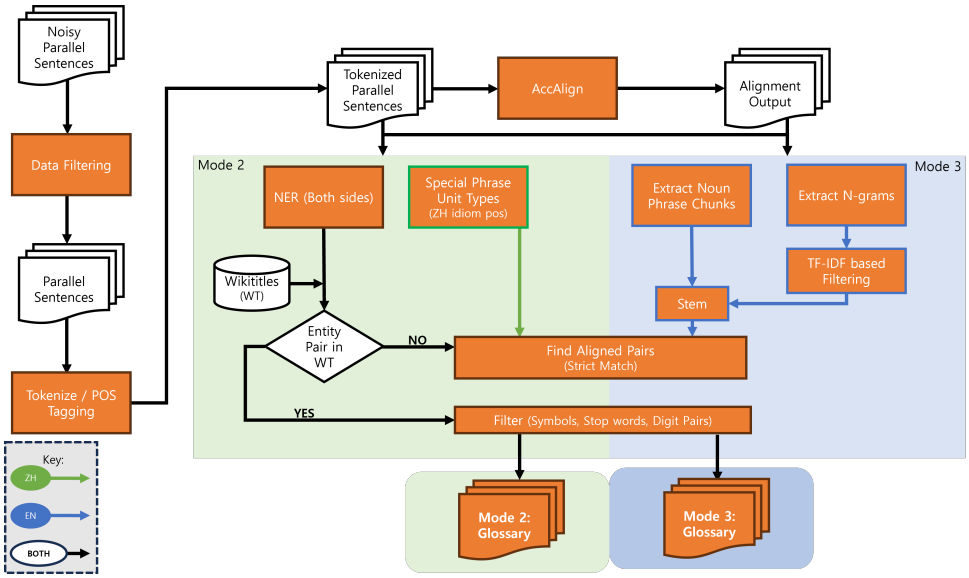


Figure 1: Data Construction Process

processed from the English side when generating Mode 3 candidates. We extracted n-grams ($1 \leq n \leq 4$) with high TF-IDF scores (Sparck Jones, 1988), as well as noun phrases (Loria, 2018), as the Mode 3 candidates. We take all Mode 3 candidate pairs that contain an appropriately aligned terminology, stem the English terms using `nltk` (Bird et al., 2009), and then randomly select of a maximum of ten term pairs for the Mode 3 glossary.

3.5 Development Data

The official Chinese-English development data is relatively small, thus we supplement it with a subset of the allowed data. We construct a supplemental development data with a random proportional sample from each provided training data source, which consists of 1,000 identical sentences throughout the three modes. Furthermore, we construct terminology for the different modes according to the above mentioned process. Additionally, we filter stop words and terms in neither the source nor target texts.

4 Models

This section presents two distinct models designed to incorporate terminologies into NMT models. The first model, Terminology Self-selection Neural Machine Translation (TSSNMT), employs a shared encoder architecture featuring a gating mechanism. This mechanism empowers the model to make decisions regarding the proportion of the source sentence and the terminologies to be processed during generation. The second model, ForceGen Transformer (ForceGen-T), takes a more straightforward approach, utilizing a standard Transformer model with force decoding (Reheman et al., 2023) and copy mechanism (Song et al., 2019). This approach enforces the model to initially generate the predefined terminologies before generating the remainder of the sentence. Copy mechanism is applied to replicate source-side target terminologies in the output.

4.1 TSSNMT

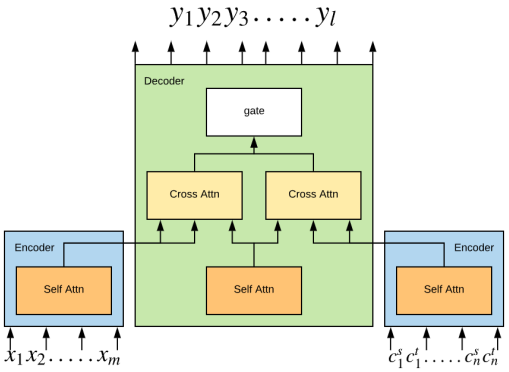


Figure 2: TSSNMT Model Structure. x, y, and c denote source, target and corresponding terminology respectively.

We have implemented the TSSNMT model with minor changes to the transformer architecture. The model has two encoders, as shown in Figure 2. Each encoder receives input in source sentences and source-target pair terminologies. These two en-

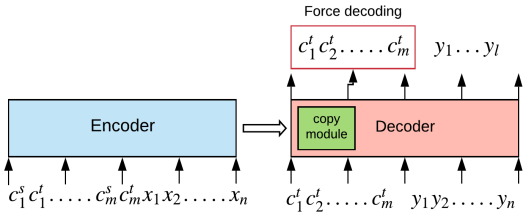


Figure 3: ForceGen model structure. x, y, and c denote source, target and corresponding terminology respectively.

coders share parameters and encode both the source sentences and the terminologies. The decoder calculates cross-attentions with these two encoder hidden states separately and then projects them to a gating mechanism (Bapna and Firat, 2019). Sharing parameters allows the model to decide weight distribution across the source and the terminology during generation.

4.2 ForceGen Transformer

We tailor a Transformer-based model to ensure the appearance of given terminologies in the generated output. Modifying the input format and decoding process incorporates copy mechanism on the source side, allowing it to copy the target terminology from the provided terminology pairs. Table 2 refers to the input sentence for this purpose.

During decoding, the model is reinforced to generate the given terminologies in a teacher-forcing manner. This approach aligns with findings by Reheman et al. (2023), who force a model to generate Translation Memory (TM) to enhance model performance. Instead, we provide the model with the terminologies, expecting it to consider them attentively during the decoding process. Once the terminologies are successfully generated, the model decodes the remainder of the input text. Copy mechanism enforces the source-side target constraints into the output. This approach is inline with that of Song et al. (2019), where copy mechanism significantly improves the ratio of terminology occurrences in the output.

We conduct preliminary experiments by training our models using the IWSLT17 (Cettolo et al., 2017) Chinese-English data and MUSE dictionary (Conneau et al., 2017) to assess the impact of the copy mechanism and force decoding. The primary objective is to determine whether the copy mechanism and the force decoding technique could complement each other. The outcomes, as presented in Table 3, reveal that the model yields the most favorable results when both the copy mechanism and force decoding are concurrently applied. This finding underscores the benefit of replicating source-side target terminology during the generation process, as it significantly aids in generating pre-specified terminologies during the force decoding phase. Consequently, when generating output after force decoding, the model effectively focuses on the target terminologies generated during decoding, facilitating successful incorporation of these terminologies into the final production. We apply both methods to our model, ForceGen-T.

5 Experiments

5.1 Evaluation setting

5.1.1 Pseudo test data

Both the provided WMT Terminology test data and blind data for the Chinese-English language pair contain only Chinese source lines and no target, so to evaluate the model, we constructed artificial target answers using ChatGPT (Ouyang et al., 2022) and reviewed the produced data manually. We then use this data as the test data to evaluate the model.

5.1.2 Evaluation metrics

The evaluation criteria for this translation task include overall terminology translation, terminology usage, and translation quality. We chose SacreBLEU (Post, 2018) , COMET (Rei et al., 2020), chrF (Popović, 2015) and Copy Success Rate (CSR). SacreBLEU and COMET scores are commonly used metrics in machine translation quality. For terminology translation and usage, we utilize CSR, which we define as the appearance rate of the desired terminology in the inferred text.

5.2 Experimental details

We use sentencepiece (Kudo and Richardson, 2018) to learn a joint byte pair encoding with a vocabulary size of 32K. Our preprocessing strategy involves pre-tokenizing Chinese data through an in-house Chinese tokenizer, while English data is exclusively tokenized using the Sentencepiece model. Please note that the training data tokenization process slightly differs from the data construction described in Section 3.

For all the experiments, we build upon the scale of the Transformer Big model (Vaswani et al.,

Source	郝仁, 人如其名, 是一个好人。
Term	{“en”: “Hao Ren”, “zh”: “郝仁” }
Modified source	郝仁<C> Hao Ren </C> 郝仁, 人如其名, 是一个好人
Modified target	Hao Ren </C> Hao Ren, as his name suggests, is a good man

Table 2: ForceGen training data sample.
<C>, </C> are the separation token that distinguishes the source sentence from the terminologies.

Model	COMET	SacreBLEU	chrF	CSR
Baseline	0.7274	18.78	42.09	78%
+Copy	0.7347	19.44	42.16	92%
+Force decoding	**0.7371**	**20.10**	**43.31**	**94%**

Table 3: Preliminary experiment results of ForceGen-T trained with IWSLT17 Chinese-English data.

Test data	Model	COMET	SacreBLEU	chrF	CSR
Test data	Baseline	0.6932	17.13	45.13	54.35%
	TSSNMT	0.7205	**23.04**	48.68	**75.86%**
	ForceGen-T	**0.7380**	22.02	**51.00**	73.71%
Blind data	Baseline	0.6918	16.55	45.88	65.07%
	TSSNMT	0.7181	**23.26**	49.18	83.45%
	ForceGen-T	**0.7336**	20.96	**51.57**	**89.38%**

Table 4: Experiment results. Please note that the scores are measured with Chat-GPT generated references.

2017) architecture implemented using our proprietary toolkit. This model consists of 12 encoder layers and 6 decoder layers, providing a strong foundation for effectively integrating specified terminologies into the output. The specific configuration of each approach varies according to the respective model specifications. We list detailed configurations in Appendix B.

6 Results

Table 4 shows the Chinese-English translation results on the WMT'23 Terminology Task. We compare two approaches - TSSNMT and ForceGen-T against the baseline Transformer Big model. Both TSSNMT and ForceGen-T significantly outperform the baseline model in all automatic evaluation metrics. Highly elevated CSR scores underscore the successful integration of provided terminologies into the translated output. In contrast, higher scores in various syntactic and semantic metrics (COMET, SacreBLEU, and chrF) indicate the fluency and adequacy of the generated translations. Within the test data, both TSSNMT and ForceGen-T exhibit similar performance levels. However, when evaluating based on the CSR score, TSSNMT surpasses ForceGen-T by approximately 2%p. In contrast, within the blind data, ForceGen-T consistently demonstrates superior scores compared to TSSNMT, with particularly notable advantages in CSR scores.

7 Conclusion

This paper presents the comprehensive procedure of our submissions for the WMT'23 Terminology Shared Task. Our approach involves meticulous refinement and pre-processing of the provided data, subsequently used to train our models. We investigate and implement two strategies for effectively integrating the given terminologies into the output, demonstrating their superior performance compared to the baseline. The result shows that our approach can significantly improve translation accuracy by increasing the recall of terminologies. As a future endeavor, we aim to extend the validation of our approach to other languages.

8 Limitations

In this paper, we propose two successful terminology integration approaches in NMT. We confirm that our models achieve significant performance gains over the baseline model. Still, it is essential to note that these observed improvements are specific to a particular language pair, Chinese to English. Therefore, further experiments on a wide range of language pairs, including those with morphologically complex structures, are needed to validate the broader efficacy of our approaches.

It is worth noting that the inference speed of ForceGen-T linearly correlates with the number of terminologies that need to be generated. ForceGen-T is forced to generate the given terminologies first in decoding, inevitably requiring additional inference time. Consequently, the inference speed of ForceGen-T is slower than that of the baseline Transformer model.

References

Ankur Bapna and Orhan Firat. 2019. Non-parametric adaptation for neural machine translation. *arXiv preprint arXiv:1903.00058*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 2–14, Tokyo, Japan. International Workshop on Spoken Language Translation.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. *arXiv preprint arXiv:1906.01105*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. *arXiv preprint arXiv:1704.07138*.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Steven Loria. 2018. textblob documentation. *Release 0.15*, 2.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Ruixuan Luo, Jingjing Xu, Yi Zhang, Zhiyuan Zhang, Xuancheng Ren, and Xu Sun. 2019. Pkuseg: A toolkit for multi-domain chinese word segmentation. *CoRR*, abs/1906.11455.

Alexander Molchanov, Vladislav Kovalenko, and Fedor Bykov. 2021. PROMT systems for WMT21 terminology translation task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 835–841, Online. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. *arXiv preprint arXiv:1804.06609*.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Abudurexiti Reheman, Tao Zhou, Yingfeng Luo, Di Yang, Tong Xiao, and Jingbo Zhu. 2023. Prompting neural machine translation with translation memories. *arXiv preprint arXiv:2301.05380*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019. Code-switching for enhancing nmt with pre-specified translation. *arXiv preprint arXiv:1904.09107*.

Karen Sparck Jones. 1988. *A Statistical Interpretation of Term Specificity and Its Application in Retrieval*, page 132–142. Taylor Graham Publishing, GBR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ke Wang, Shuqin Gu, Boxing Chen, Yu Zhao, Weihua Luo, and Yuqi Zhang. 2021. TermMind: Alibaba's WMT21 machine translation using terminologies task submission. In *Proceedings of the Sixth Conference on Machine Translation*, pages 851–856, Online. Association for Computational Linguistics.

Weikang Wang, Guanhua Chen, Hanqing Wang, Yue Han, and Yun Chen. 2022. Multilingual sentence transformer as a multilingual word aligner. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2952–2963, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Appendix

Data	Num of raw data	Num of filtered data	ratio
Back-translated news	16,943,688	16,349,073	96.49%
CCMT Corpus(casia2015)	1,048,400	1,046,410	99.81%
CCMT Corpus(casict2011)	1,512,478	952,259	62.96%
CCMT Corpus(casict2015)	2,019,011	1,968,537	97.50%
CCMT Corpus(datum2017)	718,025	656,980	91.50%
CCMT Corpus(neu2017)	1,967,605	1,894,805	96.30%
News Commentary v18.1	311,904	309,410	99.20%
ParaCrawl v9	10,508,286	6,599,206	62.80%
UN Parallel Corpus v1.0	12,354,729	12,206,477	98.80%
WikiMatrix	2,276,736	1,035,916	45.50%
Total	**49,660,862**	**43,019,073**	**86.63%**

A.1: Data Filtering

Train data	COMET	sacreBLEU	chrF	CSR
raw data	0.5829	12.36	34.41	72.54%
filtered data	0.6445	15.04	40.58	73.64%

A.2: Comparison of WMT'23 raw data and filtered data on the test data

B Appendix

Training configuration	Hyper-parameters
embedding size	1024
num of encoder layers	12
num of decoder layers	6
num of heads	16
hidden size	1024
bottleneck size	4096
dropout rate	0.15
optimizer	fusedadam
learning rate	1.8
lr scheduler	noam
warm up step	4000
strategy	deepspeed_stage_2(Rajbhandari et al., 2020)

B.1: Training Configure

HW-TSC's Participation in the WMT 2023 Automatic Post Editing Shared Task

Jiawei Yu[1]*, Min Zhang[2], Yanqing Zhao[2], Xiaofeng Zhao[2], Yuang Li[2], Chang Su[2], Yinglu Li[2], Miaomiao Ma[2], Shimin Tao[2], Hao Yang[2]
[1]School of Informatics, Xiamen University, China
[2]Huawei Translation Services Center, Beijing, China
yujiawei@stu.xmu.edu.cn
{zhangmin186,zhaoyanqing,zhaoxiaofeng14,liyuang3,suchang8,liyinglu,
mamiaomiao, taoshimin, yanghao30}@huawei.com

Abstract

The paper presents the submission by HW-TSC in the WMT 2023 Automatic Post Editing (APE) shared task for the English-Marathi (En-Mr) language pair. Our method encompasses several key steps. First, we pre-train an APE model by utilizing synthetic APE data provided by the official task organizers. Then, we fine-tune the model by employing real APE data. For data augmentation, we incorporate candidate translations obtained from an external Machine Translation (MT) system. Furthermore, we integrate the En-Mr parallel corpus from the FLORES-200 dataset into our training data. To address the overfitting issue, we employ R-Drop during the training phase. Given that APE systems tend to exhibit a tendency of 'over-correction', we employ a sentence-level Quality Estimation (QE) system to select the final output, deciding between the original translation and the corresponding output generated by the APE model. Our experiments demonstrate that pre-trained APE models are effective when being fine-tuned with the APE corpus of a limited size, and the performance can be further improved with external MT augmentation. Our approach improves the TER and BLEU scores on the development set by -2.42 and +3.76 points, respectively.

1 Introduction

Automatic Post-Editing (APE) is a post-processing task in a Machine Translation (MT) workflow, aiming to automatically identify and correct errors in MT outputs (Chatterjee et al., 2020a). WMT has been holding APE task competitions in different languages and fields since 2015. Similar to WMT 2022, WMT 2023's APE task still focuses on the En-Mr language pair. Participants are provided with a training set comprising 18,000 instances, a development set, and a test set, with each containing 1,000 instances. Each dataset consists of triplets — the source (*src*) sentences, the corresponding machine-translation (*mt*) outputs, and the human post-edited versions (*pe*) of the translations. In this task, the source sentences have been translated into the target language by using a state-of-the-art neural MT system to get the machine-translation data. The provided data encompasses diverse domains, such as healthcare, tourism, and general/news. In addition, the synthetic training data is offered to participants, which is created by taking a parallel corpus, where the source data is translated using an MT system, and the references are considered as post-edits. Furthermore, participants are permitted to utilize any additional data for systems training.

Typically, training an APE model requires large amount of training data. However, obtaining *pe* is an expensive task in terms of time and money. As a result, there exists a scarcity of large-scale APE datasets.

To address this challenge, numerous data augmentation techniques have been proposed (Junczys-Dowmunt and Grundkiewicz, 2016; Negri et al., 2018; Lee et al., 2020; Wei et al., 2020; Zhang et al., 2023). Wei et al. (2020) augment the APE training data with translations generated using a different MT system. Huang et al. (2022) train an external MT to obtain more datasets consistent with APE tasks. They also use Google translation to back translate the post-edits in the training set. Deoghare and Bhattacharyya (2022) augment the APE data by generating phrase-level APE triplets using SMT phrase tables. To ensure the quality of the synthetic data, they employe the LaBSE technique (Feng et al., 2022) to filter low-quality triplets.

In our method, we use Google translation to back translate the post-edits in the training set. Subsequently, our dataset is structured as follows: the concatenation of source sentence, back translation and machine translation as the input, while the

*Work done during internship at Huawei

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 926–930
December 6–7, 2023.

post-edits serve as the reference output. Additionally, we incorporate En-Mr parallel sentences from FLORES-200 (Costa-jussà et al., 2022) dev and test data to our training set. Given that we have an En-Mr parallel corpus only and lack machine translation data, we directly utilize English sentences as the source input and Marathi sentences as the post-edits. Furthermore, we use R-Drop (Liang et al., 2021), which regularizes the training inconsistency induced by dropout and has been proven beneficial for different kinds of models.

Chatterjee et al. (2020b) have proven that APE systems often make unnecessary edits to translation output. To mitigate this issue of over-correction, we employ a sentence-level QE system to determine the final output, selecting between the APE system's output and the original machine-translated (*mt*) version.

When being evaluated on the development set, our approach improves the TER (Snover et al., 2006) by -2.42 points and the BLEU score (Papineni et al., 2002) by +3.76 points.

The contributions of our work are as follows:

- We employ two approaches for data augmentation: (1) We utilize Google translation to back translate the post-edits to get *src'*. (2) We add English and Marathi data from the FLORES-200 dataset to our training set.
- We utilize R-Drop to address over-fitting concerns and enhance the generalization capabilities of our model.
- We employ a sentence-level QE system to select the most appropriate output, choosing between the APE-generated output and the original translation.

2 Related Work

Last year's WMT22 APE shared task mainly focuses on transfer learning and data augmentation. Huang et al. (2022) employ the existing data to train an En-Mr translation model as a data augmentation method. Additionally, they utilize an external MT system to generate back-translations, which can be used to add a set of parallel corpora for the model to learn the rules of post-edits. Adapters are also incorporated into the APE model, allowing the training data to be steered to different adapters based on the output of a trained classifier. This facilitates the model in learning post-editing rules specific to different translations.

Deoghare and Bhattacharyya (2022) use two separate encoders to generate representations for src and mt. They also employ a pre-trained language model to initialize the weights for both our encoders. For data augmentation, they leverage external MT candidates and generate phrase-level APE triplets using SMT phrase tables. Furthermore, they filter low-quality APE triplets from the synthetic data using LaBSE-based filtering. They also use a sentence-level QE system to select the final output between the APE-generated output and the original translation.

With experience in previous competitions, we also utilize an external MT system to generate back-translations. Additionally, we adopt a sentence-level QE system for selecting the final output.

3 Dataset

3.1 Data source

We use the WMT22 official En-Mr APE dataset, which consists of a training set and a development set. The training set consists of 18,000 APE triplets across domains, such as healthcare, tourism, and general/news. We first use synthetic data with 2.57M instances to pre-train our model, which was prepared as a part of the 2022 APE shared task. Furthermore, we enrich our training set by incorporating 2,000 En-Mr parallel sentences from the FLORES-200 dataset. FLORES-200 is a high quality, many-to-many benchmark dataset, which contains about 204 languages. In our approach, we specifically extract the English and Marathi parallel corpus from this dataset for training purposes.

4 Model

Figure 1 shows the architecture of our APE model. In this section, we provide the details of our approach.

4.1 Fine-tuned Transformer

We basically treat the APE task as an NMT-like problem, which takes *src* and *mt* as input and generates *pe* autoregressively. Following previous works, we use a special token <*s*> to concatenate src and mt to generate the input sentence: [*src,* <*s*>*, mt*], while the target sentence is *pe*. Initially, we pre-train the APE model using the standard Transformer (Vaswani et al., 2017) structure on 2.57M synthetic training data. However, since there is a mismatch between the synthetic data and the real data in our task, we further fine-tune the APE

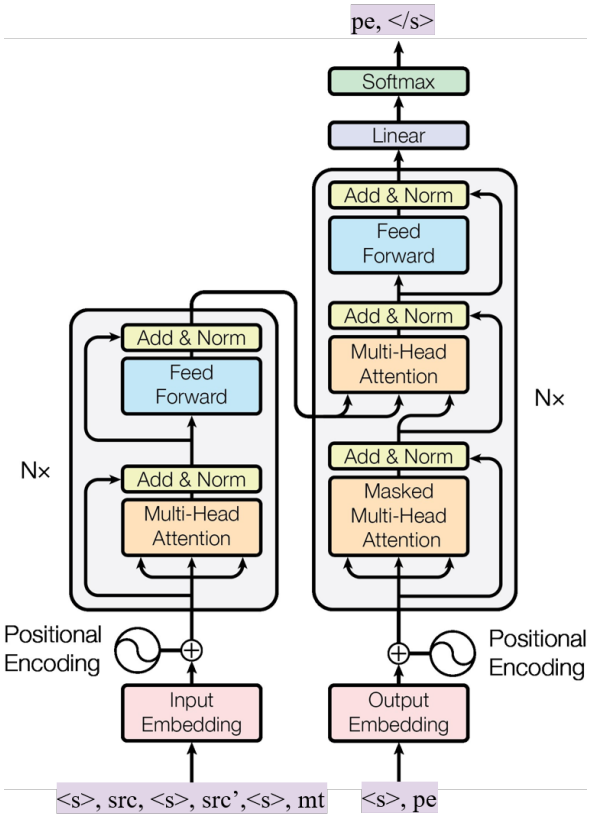


Figure 1: This figure shows the architecture of our model, where mt and augmented src' are concatenated with src before being input into the encoder, and post-edits are generated with the decoder.

model using the APE dataset. To further solve the problem of data scarcity, following (Huang et al., 2022), we use the Google translation system to create the *src'* from the provided *pe* text. We simply concatenate the *src'* with the original *src* and *mt* to form the new input: [*src, <s>, src', <s>, mt*]. Then, we use it in the same way as before, aiming to have the model learn complementary information from *src* and *src'*. During inference, the same input [*src, <s>, src', <s>, mt*] is employed to generate the output, thereby enabling the utilization of the external information derived from *src'*.

We also employ R-Drop during the fine-tuning stage to mitigate overfitting and enhance the generalization capabilities of our model.

4.2 Sentence-Level Quality Estimation

We use wmt22-cometkiwi-da (Rei et al., 2022) as our sentence-level QE model, which is a COMET quality estimation model. This model can be used for reference-free MT evaluation. It receives a source sentence and the respective translation and returns a single score between 0 and 1 that reflects the quality of the translation, where 1 represents a perfect translation. We use this model to rate both the original machine translation and the output generated by our APE system. We then compare the ratings for both sequences and select the one with a higher rating as the final output.

5 Experiment

5.1 Settings

Our model is implemented with fairseq (Ott et al., 2019). Note that the vocabulary and encoder/decoder embeddings of our model are shared between two languages and contain 30K subtokens. All models are trained on a Nvidia Tesla V100 GPU with 32GB memory. We use the batch size of 30,720 tokens in the pre-training stage and 8,192 tokens in the fine-tuning stage. We leverage the FP16 (mixed precision) training technique to accelerate the training process. In all stages, we apply the Adam optimizer(Kingma and Ba, 2015) with β_1 = 0.9, β_2 = 0.98 to train the model, where the inverse square root schedule algorithm and warmup strategy are adopted for the learning rate. Concretely, We use a learning rate of 5e-4 with 20k warm-up steps in the pre-training stage and a learning rate of 5e-5 with 4k warm-up steps in the fine-tuning stage. Besides, we set the dropout to 0.1 in the pre-training stage, 0.3 in the fine-tuning stage, and the value of label smoothing to 0.1 in all stages. Early stopping is adopted with patience 10 and 30 epochs during pre-training and fine-tuning, respectively. During inference, we use beam search with a beam size of 10. Finally, we employ BLEU to evaluate the model performance. TER and newly added evaluation metric chrF (Popovic, 2015) are also used to evaluate the model output.

System	BLEU↑	TER↓
Baseline (Do nothing)	64.62	22.93
+APE Data Fine-tuning	66.20	22.82
+External MT	66.46	22.12
+Flores data	66.83	22.01
+R-Drop	67.76	21.12
+Sentence-level QE	**68.38**	**20.51**

Table 1: Results on the WMT23 APE development set. A situation with a higher BLEU score but lower TER indicates a better result.

5.2 Result

Table 1 shows the experimental results evaluated on the dev set, where the baseline result is produced by directly calculating scores between the provided MT and PE.

The first experiment is performed by fine-tuning all parameters of the pre-trained Transformer on the official training set, which obtains 2+ performance gains compared with the baseline. This demonstrates that fine-tuning the pre-trained NMT model on the limited dataset can be useful. The experiment of adding external MT for data augmentation shows significant improvements in performance. The third row in Table 1 shows the results of the experiment where we add FLORES-200 data. In the fourth row, we show the results when R-Drop is adopted in our training stage. Toward the end, we utilize a sentence-level QE system to rate both the original translation and the APE output. We then select one of them with a higher rating as the final output of our APE system. With the combination of the APE model and sentence-level QE system, we see that the TER decreases to 20.50, and the BLEU score increases to 68.38 points.

6 Conclusion

This paper presents our APE system submitted to the WMT 2023 APE English-Marathi shared task. In our approach, we initially employ the data augmentation method to build the [*src,* <*s*>*, src',* <*s*>*, mt*] additional training datasets. We augment our training data by incorporating the En-Mr parallel sentences from Flores-200 dataset. We mitigate overfitting by employing R-Drop during the training phase. Moreover, we explore the sentence-level QE system to discard low-quality APE outputs. Evaluation of our APE system shows that our approach achieves significant gains on the WMT-22 APE development sets.

Limitations

One limitation of our approach is that while we utilize a sentence-level QE system to assess the quality of the APE output and the original translation, the APE system itself does not directly benefit from this evaluation process. While the QE system helps us identify and discard poor-quality APE outputs, it does not contribute to the improvement of the APE system itself.

Acknowledgements

We would like to thank the anonymous reviewers. Their insightful comments helped us in improving the current version of the paper.

References

Rajen Chatterjee, Markus Freitag, Matteo Negri, and Marco Turchi. 2020a. Findings of the WMT 2020 shared task on automatic post-editing. In *Proc. of WMT@EMNLP*.

Rajen Chatterjee, Markus Freitag, Matteo Negri, and Marco Turchi. 2020b. Findings of the WMT 2020 shared task on automatic post-editing. In *Proc. of WMT@EMNLP*.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*.

Sourabh Dattatray Deoghare and Pushpak Bhattacharyya. 2022. IIT bombay's WMT22 automatic post-editing shared task submission. In *Proc. of WMT*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proc. of ACL*.

Xiaoying Huang, Xingrui Lou, Fan Zhang, and Tu Mei. 2022. Lul's WMT22 automatic post-editing shared task submission. In *Proc. of WMT*.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Log-linear combinations of monolingual and bilingual neural machine translation models for automatic post-editing. In *Proc. of WMT*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proc. of ICLR*.

WonKee Lee, Jaehun Shin, Baikjin Jung, Jihyung Lee, and Jong-Hyeok Lee. 2020. Noising scheme for data augmentation in automatic post-editing. In *Proc. of WMT@EMNLP*.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. 2021. R-drop: Regularized dropout for neural networks. In *Proc. of NeurIPS*.

Matteo Negri, Marco Turchi, Rajen Chatterjee, and Nicola Bertoldi. 2018. ESCAPE: a large-scale synthetic corpus for automatic post-editing. In *Proc. of LREC*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proc. of NAACL-HLT*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proc. of ACL*.

Maja Popovic. 2015. chrf: character n-gram f-score for automatic MT evaluation. In *Proc. of WMT@EMNLP*.

Ricardo Rei, Marcos V. Treviso, Nuno Miguel Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte M. Alves, Luísa Coheur, Alon Lavie, and André F. T. Martins. 2022. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. In *Proc. of WMT*.

Matthew G. Snover, Bonnie J. Dorr, Richard M. Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proc. of AMTA*. Association for Machine Translation in the Americas.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. of NeurIPS*.

Daimeng Wei, Hengchao Shang, Zhanglin Wu, Zhengzhe Yu, Liangyou Li, Jiaxin Guo, Minghan Wang, Hao Yang, Lizhi Lei, Ying Qin, and Shiliang Sun. 2020. Hw-tsc's participation in the WMT 2020 news translation shared task. In *Proc. of WMT@EMNLP*.

Min Zhang, Xiaofeng Zhao, Zhao Yanqing, Hao Yang, Xiaosong Qiao, Junhao Zhu, Wenbing Ma, Su Chang, Yilun Liu, Yinglu Li, Minghan Wang, Song Peng, Shimin Tao, and Yanfei Jiang. 2023. Leveraging chatgpt and multilingual knowledge graph for automatic post-editing. In *International Conference on Human-Informed Translation and Interpreting Technology (HiT-IT 2023)*. Accepted for publication.

Neural Machine Translation for English - Manipuri and English - Assamese

Goutam Agrawal and **Rituraj Das** and **Anupam Biswas** and **Dalton Meiti Thounaojam**
National Institute of Technology, Silchar
{goutam_pg_22, rituraj_pg_22, anupam, dalton}@cse.nits.ac.in

Abstract

The internet is a vast repository of valuable information available in English, but for many people who are more comfortable with their regional languages, accessing this knowledge can be a challenge. Manually translating this kind of text, is a laborious, expensive, and time-consuming operation. This makes machine translation an effective method for translating texts without the need for human intervention. One of the newest and most efficient translation methods among the current machine translation systems is neural machine translation (NMT). In this WMT23 shared task: low resource indic language translation challenge, our team named ATULYA-NITS used the NMT transformer model for the English to/from Assamese and English to/from Manipuri language translation. Our systems achieved the BLEU score of 15.02 for English to Manipuri, 18.7 for Manipuri to English, 5.47 for English to Assamese, and 8.5 for Assamese to English.

1 Introduction

In countries like India, linguistic diversity is a significant aspect, with a multitude of languages varying across different regions. India officially recognizes 23 languages (Das et al., 2020) (e.g., Hindi, Sanskrit, Assamese, Odia, etc.), and alongside these, there are several hundred unofficial local languages spoken by communities. Despite India's vast population of approximately 1.4 billion, only about 11% of the population is proficient in English (Azam et al., 2013).

This language barrier becomes crucial when considering the abundance of valuable resources available on the internet, mostly in English, as a significant proportion of people in India cannot fully comprehend this content. Consequently, there arises a pressing need to translate such valuable information into local languages to facilitate knowledge sharing among the population. Such knowledge dissemination is crucial not just for business purposes but also for enabling the exchange of feelings, opinions, and actions, thereby fostering better communication and understanding among people from diverse linguistic backgrounds.

Manual translation of such copious amounts of content would be extremely laborious and time-consuming, making automatic machine translation an indispensable solution. However, machine translation for Indian languages presents its own set of challenges (Singh et al., 2021). One key challenge is the scarcity of parallel corpora, as there are fewer resources available for Indian languages compared to more widely spoken foreign languages. Moreover, the structural differences between Indian languages and English, particularly in terms of morphological richness and word order, pose significant obstacles to accurate translation. For instance, English follows a Subject-Verb-Object (SVO) word order, whereas Indian languages like Assamese and Manipuri, follow a Subject-Object-Verb (SOV) word order (Bora, 2015). Furthermore, English is a fusional language, while Assamese and Manipuri are agglutinative languages (Singh and Singh, 2022; , leading to distinct syntactic and morphological complexities that further complicate the translation process.

We participated in the Low-Resource Indic Language Translation task on translating two language pairs i.e. English to/from Assamese, and English to/from Mizo. We did the preprocessing of the given dataset and applied a neural machine translation technique i.e. transformer model. The performance was evaluated using the widely used evaluation metric BLEU. The rest of the paper is organized as follows: Section 2 discusses the existing machine translation systems and techniques tailored to Indian languages. In Section 3, we present details about the dataset, preprocessing of the dataset, and transformer model. In section 4, we discussed about the result. Finally, in Section 5, we

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 931–934

conclude with a discussion of the future prospects.

2 Literature Survey

Over the past few decades, machine translation (MT) has been the subject of extensive research. Researchers have explored various approaches in this field, including rule-based MT (Das and Baruah, 2014; Forcada et al., 2011), corpus-based MT, also known as data-driven MT (Laskar et al., 2022; Laitonjam and Singh, 2021; Singh and Bandyopadhyay, 2010), and hybrid-based MT (Laitonjam and Singh, 2022). Each of these approaches has its own advantages and disadvantages.

In rule-based MT, systems analyze the source text to create an intermediate representation, and depending on this representation, it can be further categorized into transfer-based (TBA) and interlingua-based (IBA) approaches. The corpus-based approach, on the other hand, relies on large parallel corpora consisting of text and their translations to acquire translation knowledge and is sub-divided into two sub-types, i.e. statistical machine translation (SMT) and example-based machine translation (EBMT). SMT generates translations using statistical models that combine language models and translation models with decoding algorithms. In contrast, EBMT uses existing translation examples to generate new translations. Hybrid-based machine translation combines aspects of both rule-based and corpus-based approaches to address their respective limitations.

The machine translation performance for Indian language pairs (e.g., Hindi, Bengali, Tamil, Punjabi, Gujarati, and Urdu) into English achieves an average accuracy of only 10%, (Khan et al., 2017) highlighting the need for improved machine translation systems for these languages. Neural Machine Translation (NMT) has emerged as a novel and promising technique for various languages, exhibiting remarkable results (Devi and Purkayastha, 2023; Laskar et al., 2022, 2021). In this paper, we have applied the transformer model to the English-Assamese and English-Manipuri language pair (Laskar et al., 2021; Singh and Singh, 2022)

3 Methodology and Evaluation

3.1 Dataset Details

The English-Assamese parallel corpus (Pal et al., 2023) comprised a grand total of 53,000 sentence pairs, while the Assamese monolingual corpus contained nearly 2.6 million sentences. Moving over to the English–Manipuri parallel corpus (Pal et al., 2023), it included a substantial 24,300 aligned sentence pairs. As for the Manipuri monolingual dataset, it contained roughly 2.1 million sentences.

3.2 Data Preprocessing

The dataset may contain repetition of sentences with the same source and the same target translation, sentences with the same source but different translations, sentences with different source text but the same translation. To address these issues, a solution was implemented by selecting unique sentence pairs from all available sentences and removing the duplicates. Sentences repeated more than once were completely removed to avoid ambiguity in determining the correct translation for a given source and vice versa. This preprocessing step aimed to ensure that the training and test sets did not contain the same sentences, which could result in better predictions for the test set but incorrect predictions for new sentences. Some additional preprocessing steps were carried out, including removing sentences with a length greater than 50, removing noisy translations and unwanted punctuations, filtering out sentences in other languages by applying language identification, and filtering out sentences containing HTML tags, illegal characters, and invisible characters. Finally, the dataset was split into training, testing, and validation sets, following shuffling. The English-Assamese parallel corpus was segregated into 49,500 for training, 2,000 for validation, and 1,000 for testing. Similarly, the English-Manipuri parallel corpus was divided into 21,000 for training, 2,000 for validation, and 1000 for testing.

3.3 Transformer Model

The Transformer model(Vaswani et al., 2017) is a powerful architecture used in tasks like machine translation. It excels in natural language processing, employing a technique called "self-attention" to process sequential data effectively. Unlike traditional models, it considers the context of the entire sequence, using multiple self-attention mechanisms known as "attention heads" to capture different relationships between words. Positional encoding is added to understand the word order. In machine translation, it consists of an encoder and a decoder communicating through attention mechanisms.

For the task of the English-Assamese language pair, along with the provided parallel corpus (Pal et al., 2023), we also used the monolingual corpus to create the vocabulary for the English and Assamese languages. The vocabulary extracted from the monolingual corpus generated a total of 107483 unique tokens in the Assamese language. The vocabulary size of the English language was 35487.

We used the transformer model to train the data. For the whole process, we used Google Colab and trained the model using a T4 GPU provided by Colab. We trained the model for 2000 training steps and 250 validation steps. We set the word vector size to 512 and used 6 layers of 512 hidden nodes. We set the transformer feed-forward size to 2048 and used 8 attention heads. We set the learning rate to 1 while using Adam optimization(Kingma and Ba, 2014). We used a batch size of 2048 with a dropout probability of 0.1 and used a label smoothing regularization technique to prevent overconfidence. The whole training process took around 4 hours when we used the batch size of 2048.

The vocabulary extracted from the monolingual corpus generated a total of 84072 unique tokens in the Manipuri language. For the English-Manipuri language pair, we trained the model for 1500 training steps and 150 validation steps, and all the remaining were similar to English-Assamese. The whole training process took around 3 hours.

4 Evaluation

4.1 Evaluation Metric

The Bilingual Evaluation Understudy (BLEU) score is a useful tool for determining the differences between translations produced by machines and those created by human translators (Papineni et al., 2002). This assessment method compares and aligns the number of n-grams in the translated output with the number of n-grams in the source text. In this context, a bigram comparison entails analyzing every word pair, while a unigram comparison relates to each individual token. It's significant to notice that this evaluation ignores the comparison's precise wording. This methodology is an improved version of a simple precision-based evaluation strategy.

4.2 Result

BLEU, chrf2, RIBES, and TER evaluation metrics on both language pairs are shown in Table 1.

Language Pair	BLEU	Chrf2	RIBES	TER
English-Assamese	5.47	21.66	0.21	0.5
Assamese-English	8.5	24.26	0.25	0.47
English-Manipuri	15.02	35.96	0.28	0.43
Manipuri-English	18.7	38.49	0.32	0.41

Table 1: The experimental result of language pairs on different evaluation metrics

5 Conclusion

In this paper, we applied NMT to the two most difficult language pairs (English-Assamese and English-Manipuri). We showed that the transformer model performs better for Indian languages. We achieved a fairly good BLEU score for the English-Manipuri language pair. So, this model can be used for domains such as tourism and education. Moreover, this transformer model is useful for various English-Indian language pair translations.

References

Mehtabul Azam, Aimee Chin, and Nishith Prakash. 2013. The returns to english-language skills in india. *Economic Development and Cultural Change*, 61(2):335–367.

Manas Jyoti Bora. 2015. Word order.

Aankit Das, Samarpan Guha, Pawan Kumar Singh, Ali Ahmadian, Norazak Senu, and Ram Sarkar. 2020. A hybrid meta-heuristic feature selection method for identification of indian spoken languages from audio signals. *IEEE Access*, 8:181432–181449.

Pranjal Das and Kalyanee K Baruah. 2014. Assamese to english statistical machine translation integrated with a transliteration module. *International Journal of Computer Applications*, 100(5).

Maibam Indika Devi and Bipul Syam Purkayastha. 2023. An exploratory study of smt versus nmt for the resource constraint english to manipuri translation. In *International Conference on Information and Communication Technology for Intelligent Systems*, pages 329–338. Springer.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25:127–144.

Nadeem Jadoon Khan, Waqas Anwar, and Nadir Durrani. 2017. Machine translation approaches and survey for indian languages. *arXiv preprint arXiv:1701.04290*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lenin Laitonjam and Sanasam Ranbir Singh. 2021. Manipuri-english machine translation using comparable corpus. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 78–88.

Lenin Laitonjam and Sanasam Ranbir Singh. 2022. A hybrid machine transliteration model based on multi-source encoder–decoder framework: English to manipuri. *SN Computer Science*, 3:1–18.

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, and Sivaji Bandyopadhyay. 2022. Improved neural machine translation for low-resource english–assamese pair. *Journal of Intelligent & Fuzzy Systems*, 42(5):4727–4738.

Sahinur Rahman Laskar, Partha Pakray, and Sivaji Bandyopadhyay. 2021. Neural machine translation for low resource assamese–english. In *Proceedings of the International Conference on Computing and Communication Systems: I3CS 2020, NEHU, Shillong, India*, pages 35–44. Springer.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Sandeep Kumar Dash, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, and Pankaj Kundan Dadure. 2023. Findings of the wmt 2023 shared task on low-resource indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Muskaan Singh, Ravinder Kumar, and Inderveer Chana. 2021. Machine translation systems for indian languages: review of modelling techniques, challenges, open issues and future research directions. *Archives of Computational Methods in Engineering*, 28:2165–2193.

Salam Michael Singh and Thoudam Doren Singh. 2022. Low resource machine translation of english–manipuri: A semi-supervised approach. *Expert Systems with Applications*, 209:118187.

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2010. Manipuri-english example based machine translation system. *Int. J. Comput. Linguistics Appl.*, 1(1-2):201–216.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

GUIT-NLP's submission to Shared Task: Low Resource Indic Language Translation

Mazida Akhtara Ahmed, Kuwali Talukdar, Parvez Aziz Boruah
Shikhar Kumar Sarma and **Kishore Kashyap**
Dept. of Information Technology, Gauhati University
Guwahati, Assam, India
14mazida.ahmed@gmail.com, kuwalitalukdar@gmail.com, parvezaziz70@gmail.com,
sks001@gmail.com, kb.guwahati@gmail.com

Abstract

This paper describes the submission of the GUIT-NLP team in the "Shared Task: Low Resource Indic Language Translation" focusing on three low-resource language pairs: English-Mizo, English-Khasi, and English-Assamese. The initial phase involves an in-depth exploration of Neural Machine Translation (NMT) techniques tailored to the available data. Within this investigation, various Subword Tokenization approaches, model configurations (exploring differnt hyper-parameters etc.) of the general NMT pipeline are tested to identify the most effective method. Subsequently, we address the challenge of low-resource languages by leveraging monolingual data through an innovative and systematic application of the Back Translation technique for English-Mizo. During model training, the monolingual data is progressively integrated into the original bilingual dataset, with each iteration yielding higher-quality back translations. This iterative approach significantly enhances the model's performance, resulting in a notable increase of +3.65 in BLEU scores. Further improvements of +5.59 are achieved through fine-tuning using authentic parallel data.

1 Introduction

Work on Machine Translation (MT) involving indigenous languages is on the rise to provide such languages a global existence rather than limiting its scope to regional geographical boundaries. But such a work is quite challenging owing to its typical characteristic being limited (low) resourced as NMT models are data hungry which tend to degrade with limited data input. Established methods like Back Translation (Sennrich et al., 2015b), Transfer Learning (Kim et al., 2019; Zoph et al., 2016), Multilingual Neural Translation (MNT) (Lakew et al., 2018; Ngo et al., 2020), Dual Learning (He et al., 2016; Wang et al., 2018) and such do exist to tackle the low-resource challenge. With the monolingual and limited parallel data provided to the teams to work with, Back Translation (BT) seemed to be an appropriate choice. In BT, a target to source model translates the target side monolingual data to generate a substantial amount of synthetic parallel data which could be augmented with the limited authentic parallel data to increase the volume of training data. Previous experiments (Sennrich et al., 2015a; Edunov et al., 2018), (Poncelas et al., 2018; Wu et al., 2019) have shown improved results in such scenarios.

The general NMT pipeline comprises of various stages like tokenization, subword tokenization, NMT model training, inference and post-editing. It should be noted that several methods are available for every stage making it difficult for the researcher to select the one that would suit the data best as each method has its own influence on the model performance. We, therefore, perform an initial investigation on two popular subword tokenization methods to find the best choice. The rest of the paper is organized as follows: Section 2 describes the methods applied for the task, Section 3 presents the experimental setup and the results obtained for the three language pairs: English-Mizo, English-Khasi and English-Assamese. Section 4 concludes the paper.

2 Methodology

The following section describes the methodology used for the task for each of the language pairs.

2.1 Data Exploration

In this section, we delve into the data used for the task, which encompasses two primary categories:

1. *Parallel Data:* This data category consists of two distinct, non-overlapping sets, specifically the training and validation set.

2. *Monolingual Data:* This category encompasses an extensive corpus with monolingual

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 935–940
December 6–7, 2023.

Language Pair	Train Set	Dev. Set
English-Mizo	50,000	1,500
English-Khasi	24,000	1,000
English-Assamese	50,000	2,000

Table 1: Parallel Data Statistics.

Language	Sentences*(in millions)*
Mizo	1.9
Khasi	0.18
Assamese	2.6

Table 2: Monolingual Data Statistics.

sentences. It is imperative to underscore that all participating teams are expressly instructed to rely exclusively on the provided data, refraining from any utilization of external resources.

Upon conducting a preliminary manual analysis of the data, several noteworthy observations have come to light:

(i) Instances exist within the corpus wherein sentences commence with multiple spaces.

(ii) Instances within the corpus also manifest where multiple spaces occur between words.

(iii) The corpus exhibits a mixture of both tokenized and untokenized sentences.

After having these disparities removed from the data, the sets are tokenized with Moses (Koehn et al., 2007) tokenizer for English, Mizo and Khasi as they share the same Roman script while Assamese is tokenized with IndicNLP[1]. Prior to tokenization of the English, Mizo and Khasi text, all characters are normalized to lowercase for consistency. With no difference in case for Assamese, this step is not required for the language. Additionally, a fundamental filtering routine is applied as part of the data preprocessing process as described below:

(a) *Removal of Empty Lines:* (Source, target) pairs containing empty lines on either the source or target side are systematically eliminated from the dataset.

(b) *Elimination of Duplicate Lines:* (Source, target) pairs characterized by duplicate lines in both the source and target segments are systematically removed. Duplicate content can introduce redundancy and skew the training process, hence necessitating their exclusion.

(c) *Relative Length-Based Filtering:* To maintain a balanced and coherent dataset, pairs where the length of the target sentence significantly exceeds that of the source sentence (or vice versa), exceeding a predetermined threshold (typically set at twice the length), are judiciously omitted.

2.2 Subword Tokenization

In the context of developing NMT models for low-resource Indian languages, subword tokenization emerges as a critical technique as it addresses out-of-vocabulary (OOV) challenge, morphological richness, facilitates cross-lingual transfer of knowledge, reduces the vocabulary size substantially. Two popular schemes are explored namely:

1. *Byte Pair Encoding (BPE):* BPE (Sennrich et al., 2015c) is a data compression technique designed to systematically merge the most common pair of character sequences. Consequently, frequent substrings are unified into single symbols, while rare words are segmented into smaller constituents. BPE is experimented in two forms:
   (a) Independent Vocabulary: This involves creating separate and independent subword vocabularies for both the source and target languages.
   (b) Shared Vocabulary: When dealing with closely related languages a shared subword vocabulary is a popular choice as it aligns (sub)words from source and target sentences into the same embedding space so as to strengthen the semantic correlation between them.

2. *Sentencepiece* (Kudo and Richardson, 2018): Though Sentencepiece (SP) has the capability to directly train subword models from raw text, eliminating the need for prior tokenization, we pre-tokenize it as (Kudo and Richardson, 2018) has shown better results with tokenized input. Also, SP supports subword regularization, which dynamically enhances the training data with on-the-fly tokenization during NMT model training. This process

[1] https://anoopkunchukuttan.github.io/indic_nlp_library/

contributes to the construction of a robust and accurate model, and it is not tied to any specific architectural configuration. Our experimentation with SentencePiece, implemented with independent vocabularies for English and Mizo, involves two main approaches:

(a) *With subword regularization:* With this method the model encounters different variations of subword splitting of the same word which could in turn be beneficial in producing a robust model for agglutinative languages. We have set the number of nbest candidates to 16 and the smoothing parameter to 0.1.

(b) *Without subword regularization.*

2.3 Using Monolingual Data

A relatively large monolingual data have been provided which could be made to use in various ways like constructing embeddings or for data augmentation. Back translation (Sennrich et al., 2015a) is a popular data augmentation method that exploits the target side monolingual data to create synthetic parallel corpus. Back Translation uses a base target→source model (initially trained on the limited genuine bitext) to translate the target side monolingual data. The synthetic data thus generated can serve as supplementary resource and could be explored in various ways.

Re-training the model on the manifold synthetic data is expected to boost up the model producing better translations. Two obvious assumptions can be made on the performance of an NMT model for low-resource scenario:

1. Data augmentation could boost up the model.
2. Also, more error-free the training data is, better is its performance.

Based on these assumptions and inspired by the previous reports on back-translation with iterations such as (Cotterell and Kreutzer, 2018; Hoang et al., 2018), we use an innovative twist to improvise model by using back translated data iteratively rather than using all in one go. In every iteration, the model is trained with increased data back translated by the previous iterations's improved model along with the original bitext thereby producing better translations for the next iteration. As the synthetic data is prone to error which could in turn hamper model performance (Poncelas et al., 2018), we add the back translated data proportionate to the size of the genuine bitext. Also, the trained model is followed by finetuning on the genuine bitext for further improvement (Tonja et al., 2023). Our method could be summarized by the following algorithm:

Algorithm 1 An innovative usage of Back Translation

Require: Authentic parallel corpus(S_0, T_0), target monolingual corpus(M), number of splits (n)
$M_0 \leftarrow Train_{(Target \rightarrow Source)}(S_0, T_0)$
$C_1, C_2, ..., C_n \leftarrow Split(M, n)$
such that $|C_i| \propto |S_0|$
$i \leftarrow 1$
while $i \leq n$ **do**

$$(S_i, T_i) = (S_0, T_0) \bigcup (M_{i-1}(\sum_{1}^{i} C_i), \sum_{1}^{i} C_i)$$

$$M_i \leftarrow Train_{(Target \rightarrow Source)}(S_i, T_i)$$

$$M_i \leftarrow FinetuneM_i(S_0, T_0)$$

$i \leftarrow i + 1$
end while

2.4 Post-Editing

The predicted translations (for English, Mizo and Khasi) have been post-edited in the following ways:

1. *Truecasing:* A truecaser model has been trained on the training set with the Moses' truecaser script.
2. Capitalizing the first character of every prediction.
3. As the text in the test set is not completely detokenized with several punctuation markers space separated, adjustments have been made to replicate the reference translations.

3 Experiments and Results

Experimental Setup: All the experiments have been conducted on the opensource NMT toolkit, OpenNMT (Klein et al., 2017). Subword vocabulary size is kept at 8000. The Transformer (Vaswani et al., 2017) has been customized to work on the small-scale dataset by simplifying the standard model. After conducting experiments

Table 3: Experimentation setup for English-Assamese

Model	En/Dec Layers	Attention Heads	Dimensions	Batch Size
Model 1	6	8	512	512
Model 2	3	4	256	256
Model 3	6	4	256	256

En/Dec : Encoder/Decoder

Table 4: Experimentation setup for English-Khasi

Model	Batch Size	BPE Vocab Size	En/Dec Layer	Attention Heads
M_1	256	6000	3	4
M_2	512	6000	3	4

En/Dec : Encoder/Decoder

Table 5: Results obtained with various Subword mechanisms (English-Mizo).

Method	BLEU	CHRF	TER	RIBES
English → Mizo				
SP_{wo_reg}	22.63	44.93	58.07	0.75
SP_{w_reg}	23.78	48.06	58.07	0.75
BPE_{sh}	23.29	46.72	59.93	0.75
BPE_{ind}	**25.58**	**48.19**	**57.35**	**0.76**
Mizo → English				
SP_{wo_reg}	20.65	40.98	72.8	0.67
SP_{w_reg}	18.51	41.32	73.7	0.67
BPE_{sh}	18.81	40.33	73.65	0.66
BPE_{ind}	**20.95**	**41.38**	**72.43**	**0.67**

SP_{wo_reg} : SentencePiece without Subword regularization

SP_{w_reg} : SentencePiece with Subword regularization

BPE_{sh} : Byte Pair Encoding with shared vocabulary

BPE_{ind} : Byte Pair Encoding with independent vocabulary

with various parameter sets (including encoder and decoder layers, heads, embedding size, and feed-forward nodes), we have determined that the optimal configuration for English-Mizo data consists of 3 encoder and 3 decoder layers, a word vector size of 512, and 2048 nodes in the feed-forward layer. For English Assamese pair, three models have been built with varying hyperparameters and training is performed in both the directions. For English Khasi, two models have been built and trained. The model descriptions for English-Assamese and English Khasi are shown in Table 3 and Table 4 respectively. All the models are trained using the Adam optimizer with an initial learning rate of 2, incorporating Noam decay and 8,000 warm-up steps. The training process continues for 200,000 steps, with validation performed every 10,000 steps. Additionally, checkpoints are saved at 10,000-step intervals, and early stopping is implemented with a patience of 4 based on validation perplexity and accuracy.

Checkpoint Selection: Throughout training, checkpoints are saved every 10,000 steps. Among all the checkpoints generated, the model with the best validation perplexity and validation accuracy is chosen as the model for testing purposes.

3.1 Results

In Table 5 we report our results on the initial experiments using various subword tokenization schemes for English-Mizo. Our results have been evaluated by four evaluation metrics as provided by the organizers. It is clear from the results that Byte Pair Encoding using independent vocabularies works best for this data. Hence, for all the future experiments, BPE with independent vocabularies is selected as the standard format. Also, it should be noted that we have reported the results obtained with BPE (shared vocabulary) as the primary results for both En → Mizo and Mizo → En directions.

Table 6 summarizes the result obtained by our method of using proportionate back translated data which is in turn generated by the model developed in the previous iteration. The baseline scores are obtained by using 1M back translated data (translated by SP_{w_reg} model) which acheives a BLEU of value of 16.77 for En → Mz. In the 1st iteration, equal size of back translated data is added to the genuine bitext and the model is trained from scratch. It is able to achieve a BLEU score of 20.42. This shows the negative impact of adding a large size synthetic data, which is not error-free, relative to the authentic parallel data. Also, a significant improvement is noticed after fine-tuning on the given authentic data. Similar results are also seen in the 2nd iteration. The successive improvement is a successful implementation of our novel usage of back translation method.

The English-Assamese and English-Khasi experiments have been conducted using various configuration of the Transformer model as shown in Table 3 and Table 4 respectively. This is done to find the optimal model configuration for the languages. Though English and Khasi share the same script, the morphologies are completely different

Table 6: English-Mizo BLEU scores with our *novel* usage of Back Translation (BT)

Method	BT Data Size	En->Mz	Mz ->En
Baseline	1M	16.77	14.40
1st Iter.	50K	20.42	16.20
	FineTuned	26.01	20.06
2nd Iter.	100K	22.04	18.19
	FineTuned	**26.63**	**20.81**

Table 9: Results for English Khasi pair.

English → Khasi				
Model	**BLEU**	**CHRF**	**RIBES**	**TER**
M_1	10.41	33.31	0.63	71.67
M_2	10.27	32.63	0.63	70.71
Khasi → English				
M_1	8.74	30.54	0.63	79.64

and as Table 5 clearly manifests, appropriate hyperparameter values can bring about significant impact in the performance. The results obtained for English → Assamese is shown in Table 7 and Assamese → English is shown in Table 8. From both the tables, we see that Model 1 has shown the best results in both English → Assamese and Assamese → English translation. We, therefore, select the results obtained for Model 1 as the primary score. For English-Khasi, the results for model (M_1) is submitted as the primary score.

Table 7: Results for English → Assamese

Model	BLEU	CHRF	TER	RIBES
Model 1	4.89	25.16	87.21	0.46
Model 2	4.27	24.59	90.13	0.43
Model 3	3.75	22.65	93.57	0.42

Table 8: Results for Assamese → English

Model	BLEU	CHRF	TER	RIBES
Model 1	5.5	25.81	80.1	0.56
Model 2	4.7	24.96	81.53	0.55
Model 3	4.14	23.73	83.41	0.53

4 Conclusion

In this study, we have provided a comprehensive overview of our Neural Machine Translation (NMT) system developed for three language pairs: English-Assamese, English-Khasi, and English-Mizo, encompassing both translation directions. Our research delved into the intricacies of model configurations (Transformer layers, heads, batch sizes, etc.) and subword tokenization schemes (Byte Pair Encoding and SentencePiece and its variants). Through rigorous experimentation, we identified and adopted the optimal configurations for each language pair.

Challenged by the inherent scarcity of data in these low-resourced language pairs, we innovatively leveraged monolingual data to augment our translation models. We have presented a novel variation of a well-established technique for addressing the challenges of low-resourced NMT systems: *Back Translation*. This adaptation yielded remarkable results, surpassing the performance of conventional Back Translation methods by a substantial margin.

5 Limitation

We use the standard tokenization implementation (Moses) for English, Mizo and Khasi. Though Moses seems to work fine for English, certain disparities (associated with language-specific characters) are observed for Mizo and Khasi, both morphologically rich languages. Similar observations are also noted for Assamese. Using a customized tokenizer for these languages is believed to enhance the results which needs further investigation.

The dataset given was too small for Neural Machine Translation trainingespecially for Khasi. Though Back Translation is a well known method for low-resource setting, merely translating and using it as a pseudo-parallel corpus may not help as the monolingual data quality also has an impact. We have not used any mechanism to judge the quality. With our method, we iteratively use incremented back translations which is observed to boost the model. But the translation data is proportional to the original parallel corpus size which hinders leveraging fully the large monolingual corpus. We would like to explore ways to fully exploit the large availability of monolingual corpus for data augmentation or linguistic embellishments. Monolingual data usage is not explored (due to time constraint as we joined late) for the English-Assamese and English-Khasi which we plan to investigate in future. Our overall system lags in producing correct translations for long sentences. Semi-automatic post editing is utilized which needs further investigations in automatising the process.

References

Ryan Cotterell and Julia Kreutzer. 2018. Explaining and generalizing back-translation through wake-sleep. *arXiv preprint arXiv:1806.04402*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. *Advances in neural information processing systems*, 29.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based transfer learning for neural machine translation between non-English languages. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 866–876, Hong Kong, China. Association for Computational Linguistics.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pages 177–180.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Surafel M Lakew, Marcello Federico, Matteo Negri, and Marco Turchi. 2018. Multilingual neural machine translation for low-resource languages. *IJCoL. Italian Journal of Computational Linguistics*, 4(4-1):11–25.

Thi-Vinh Ngo, Phuong-Thai Nguyen, Thanh-Le Ha, Khac-Quy Dinh, and Le-Minh Nguyen. 2020. Improving multilingual neural machine translation for low-resource languages: French, english-vietnamese. *arXiv preprint arXiv:2012.08743*.

Alberto Poncelas, Dimitar Shterionov, Andy Way, G Wenniger, and Peyman Passban. 2018. Investigating backtranslation in neural machine translation.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015a. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015b. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015c. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Atnafu Lambebo Tonja, Olga Kolesnikova, Alexander Gelbukh, and Grigori Sidorov. 2023. Low-resource neural machine translation improvement using source-side monolingual data. *Applied Sciences*, 13(2):1201.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yijun Wang, Yingce Xia, Li Zhao, Jiang Bian, Tao Qin, Guiquan Liu, and Tie-Yan Liu. 2018. Dual transfer learning for neural machine translation with marginal distribution regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Lijun Wu, Yiren Wang, Yingce Xia, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. 2019. Exploiting monolingual data at scale for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4207–4216.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

NICT-AI4B's Submission to the Indic MT Shared Task in WMT 2023

Raj Dabre
NICT, Japan
raj.dabre@nict.go.jp

Jay Gala
AI4Bharat, India
jaygala24@gmail.com

Pranjal A. Chitale
AI4Bharat, IIT Madras, India
cs21s022@cse.iitm.ac.in

Abstract

In this paper, we (Team NICT-AI4B) describe the MT systems that we submit to the Indic MT task in WMT 2023. Our primary system consists of 3 stages: Joint denoising and MT training using officially approved monolingual and parallel corpora, backtranslation and, MT training on original and backtranslated parallel corpora. We observe that backtranslation leads to substantial improvements in translation quality up to 4 BLEU points. We also develop 2 contrastive systems on unconstrained settings, where the first system involves fine-tuning of IndicTrans2 Data Augmentation (DA) models on official parallel corpora and seed data used in Gala et al. (2023), and the second system involves a system combination of the primary and the aforementioned system. Overall, we manage to obtain high-quality translation systems for the 4 low-resource North-East Indian languages of focus.

1 Introduction

The increasing online presence[1] of the Indian population along with the economic growth[2] of India has necessitated the development of translation systems for Indian languages. There have been substantial efforts towards collecting monolingual and parallel corpora, as well as developing machine translation systems using them (Ramesh et al., 2022; Doddapaneni et al., 2023). Most recently, IndicTrans2 (Gala et al., 2023), an MT system, and its accompanying parallel corpus BPCC, were released. This corpus covers all 22 Indian languages covered in the 8th schedule[3] of the Constitution of India.

While IndicTrans2 has achieved comparable or better results compared to existing systems like NLLB (Costa-jussà et al., 2022), a major limitation is that there is no specific focus on language subgroups. One of such subgroups is the North-East Indian languages, which this shared task focuses on. The task focuses on translation to/from English and the following 4 North-East Indian languages: Assamese, Manipuri, Mizo and Khasi. We submit constrained as well as unconstrained MT systems for the 8 translation directions in this task. For further details on the shared task, kindly refer to (Pal et al., 2023).

We leveraged ideas such as joint multilingual denoising and MT training followed by backtranslation at scale. First, due to the small size of the official parallel corpora, we utilized available and permitted monolingual corpora for all languages involved and trained on a combination of text-infilling and MT objectives to train an initial MT system. We used this system to generate large back-translated corpora, which were combined with the official parallel corpora to train the final primary system. The back-translated corpora, due to their scale, led to improvements up to 4 BLEU as measured on the development set. We also submitted two contrastive systems: the first one was obtained via fine-tuning IndicTrans2 DA models (Gala et al., 2023) and the second one was a system combination of the primary and the aforementioned contrastive system. We observed that our first contrastive system outperformed the primary for some language pairs due to the utilization of a strong pretrained MT model as initialization and additional high-quality data being used to fine-tune them. As for our second contrastive system, we observed improvements for directions where there was a small performance gap between the primary and the first contrastive system.

2 Related Work

Our submissions leverage ideas from topics such as multilingualism (Dabre et al., 2020), denoising pre-

[1] https://datareportal.com/reports/digital-2023-india

[2] https://www.cnbc.com/2023/07/26/imf-raises-2023-economic-growth-forecast-for-india.html

[3] https://www.mha.gov.in/sites/default/files/EighthSchedule_19052017.pdf

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 941–949
December 6–7, 2023.

training (Lewis et al., 2020; Dabre et al., 2022), backtranslation (Sennrich et al., 2016), transfer learning (Zoph et al., 2016) and system combination (Heafield and Lavie, 2010, 2011).

The North-East Indian languages of focus in this shared task are all low-resource languages, and transfer learning via multilingualism is a reliable solution in this case. Transfer learning can be achieved by fine-tuning a pre-trained model (Zoph et al., 2016) but this involves two stages. On the other hand, multilingual training (Johnson et al., 2017; Dabre et al., 2020) involves implicit transfer via joint training. We explore both strategies when developing our systems.

Backtranslation (Sennrich et al., 2016) involves taking intermediate translation systems and translating monolingual corpora into another language. The synthetic-source and original target parallel corpora, can typically be orders of magnitude larger than the parallel corpora used to train the intermediate systems and when used at scale, such backtranslated corpora are known to help improve translation quality (Edunov et al., 2018) and therefore we attempt to use as much monolingual corpora as possible for backtranslation. While there are iterative backtranslation (Hoang et al., 2018) strategies where the process of model training and backtranslation is performed repeatedly, their computational complexity makes them a less attractive solution to us.

An alternative to backtranslation is denoising pretraining using monolingual corpora (Lewis et al., 2020; Dabre et al., 2022) and when combined with MT training as a joint objective (Kamboj et al., 2022) is known to significantly improve MT quality. Since backtranslation and denoising pre-training are known to be orthogonal (Liu et al., 2020), we leverage the joint denoising and MT training approach only for intermediate models which are used for backtranslation.

3 Our Systems

We submit 3 systems, one primary (constrained) and two contrastive (unconstrained).

3.1 Primary System

To create our primary system, we do the following:

1. Augment official monolingual data with the external monolingual corpora for the 4 North-East Indian languages and English.
2. Train a many-to-many encoder-decoder Transformer (Vaswani et al., 2017) model with the joint text-infilling (denoising) (Lewis et al., 2020; Dabre et al., 2022) and the MT objectives, using the augmented monolingual and official parallel corpora, respectively. To prevent the model from over-adapting to the infilling objective, we oversample the parallel corpora.
3. Use the aforementioned model to backtranslate the monolingual corpora.
4. Combine the backtranslated and official parallel corpora while oversampling the latter, and then train a many-to-many MT model.

3.2 Contrastive System #1

For our contrastive system, we investigate the potential of leveraging strong pretrained IndicTrans2 DA En-Indic and Indic-En models (Gala et al., 2023) for adaptation to newer languages and domains. It is important to note that we utilize off-the-shelf IndicTrans2 DA models trained on large-scale general-purpose corpora comprising both original and augmented backtranslated data. We refrain from using final models that are already fine-tuned with the same seed data that we use in this study, making them redundant in this context.

A trivial solution would be to adapt IndicTrans2 DA models to target languages and domains. However, this solution can often lead to catastrophic forgetting of translation ability on existing languages and domains. As a result, we explore approaches that satisfy two-fold objectives: 1) maximize the performance on a specific set of target languages and domains in the context of WMT shared task and 2) retain the overall performance on existing languages supported by the IndicTrans2 DA models. Our experiments involve a comparison of either of the approaches for adaptation of IndicTrans2 DA models to a specific set of few known and few unseen languages. We explore the following approaches:

1. **A1**: Direct fine-tuning of IndicTrans2 DA models on a combination of official parallel corpora and seed data used in Gala et al. (2023) for a set of languages under consideration for WMT Indic MT shared task.
2. **A2**: Direct fine-tuning of IndicTrans2 DA models on a combination of official parallel corpora and seed data used in Gala et al.

Parallel		Monolingual		
lang pair	# lines	lang	# lines (Org)	# lines (Aug)
as-en	50K	as	2.6M	8.05M
mz-en	50K	mz	1.9M	8.8M
kha-en	24K	kha	0.18M	0.73M
mni-en	21.6K	mni	2.14M	2.20M
		en	0	20M

Table 1: Parallel and monolingual data statistics. For the primary system, we only use the organizers' provided parallel data. We use monolingual data provided by the organizers as well as from Gala et al. (2023) and indicate the organizers' (Org) and augmented (Aug) sizes.

(2023) for all the languages supported by the IndicTrans2 DA models and WMT Indic MT shared task.

3. **A3**: Two-stage fine-tuning of IndicTrans2 DA models on 1) a combination of official parallel corpora and seed data used in Gala et al. (2023) for a set of languages under consideration for WMT Indic MT shared task, followed by 2) on a combination of official parallel corpora and seed data used in Gala et al. (2023) for all the languages supported by the IndicTrans2 DA models and WMT Indic MT shared task.

3.3 Contrastive System #2

Our second contrastive system combines the primary and first contrastive systems using a system combination approach called Multi-Engine Machine Translation (MEMT) (Heafield and Lavie, 2010, 2011). MEMT involves aligning 1-best outputs from each system using the METEOR aligner (Denkowski and Lavie, 2011), identifying candidate combinations by forming left-to-right paths through the aligned system outputs, and scoring these candidates using a battery of features. MEMT does not leverage any neural networks. We refer the readers to Heafield and Lavie (2010, 2011) for additional details.

4 Experiments

In this section, we describe the datasets, implementation and evaluation settings.

4.1 Datasets

We use the official parallel corpora and monolingual corpora provided by the organizers. We augment the monolingual corpora with those used in Gala et al. (2023). Particularly, we sample 20M English sentences, since the organizers did not provide any English monolingual data. The parallel and augmented monolingual corpora statistics are described in Table 1. For our first contrastive system, we also use a combination of BPCC seed corpora (Gala et al., 2023) and NLLB-seed corpora (Costa-jussà et al., 2022; Maillard et al., 2023) which was used in Gala et al. (2023) along with the official parallel corpora provided by the organizers for adaptation / fine-tuning IndicTrans2. Table 2 reports the statistics of different subsets used for training contrastive #1 system. For the languages primarily under consideration for the WMT Indic MT shared task, namely Assamese, Manipuri (Bengali), Khasi and Mizo, we use a total of ~196K bitext pairs encompassing seed and official parallel data.

	# langs / script	# samples
†BPCC seed	23	654,806
NLLB seed	3	18,579
WMT	4	145,321
Total	27	818,706

Table 2: Statistics of the parallel corpora used for training contrastive #1 system. † indicates that the BPCC seed also includes transliterated Sindhi (Arabic) data as released by Gala et al. (2023).

4.2 Implementation

Our primary systems are trained using YANMTT (Dabre et al., 2023). We train a single sentencepiece (Kudo and Richardson, 2018) tokenizer of 64K subwords for the Indic languages and English. We use 1M sentences per language, taken from the parallel and monolingual corpora. The model hyperparameters and optimizer details are described in Table 3. We ensure that the ratio of the official parallel and monolingual/backtranslated corpora remains balanced via temperature sampling (T=5.0) (Arivazhagan et al., 2019). We train our models till convergence with early stopping criteria with a patience of 5 and save separate checkpoints for each direction that exhibit best results for that direction based on BLEU (Papineni et al., 2002) metric on the development set. We use a fixed beam size of 4 and a length penalty of 0.8 when doing backtranslation.

For our first contrastive system, we fine-tune IndicTrans2 DA models with the standard fine-

Hyperparameter	Value
#Layers	12 (6)
Hidden size	1024 (512)
FFN hidden size	4096 (2048)
#Heads	16 (8)
Positional Encoding	Embedding
Batch size	1024 (4096)
Parameters	420M (77M)
Dropout	0.1
Label smoothing	0.1
Optimizer	Adam
#GPUs	64 (8)
GPU Type	V100
Learning rate	0.0005 (0.001)
Warmup steps	16,000
Data sampling temperature	5.0
#Train steps	∼380K (225K)

Table 3: Hyperparameter settings for primary systems. The values in round brackets, if at all, indicate those used for training smaller models, which only leverage organizers' parallel corpora.

Pair	Primary		Contrastive	
	beam	penalty	beam	penalty
as-en	16	1.4	16	1.2
kha-en	16	0.6	8	0.6
mz-en	16	1.4	16	1.4
mni-en	8	1.4	16	1.2
en-as	8	1.4	8	1.4
en-kha	16	1.4	8	1.4
en-mz	16	1.4	8	1.4
en-mni	16	1.2	8	0.8

Table 4: Optimal decoding hyperparameters settings (beam size and length penalty) obtained by performing grid search on the development set for both primary and contrastive #1 systems.

tuning hyperparameter settings following Gala et al. (2023). Our first contrastive system is based on fine-tuning of IndicTrans2 DA models (Gala et al., 2023) which uses the fairseq library[4] (Ott et al., 2019). We train our systems till convergence on the development set and use the BLEU metric for early checkpointing. Furthermore, the vocabulary of IndicTrans2 DA models (Gala et al., 2023) lacks coverage for Mizo and Khasi. To address this, we extend the vocabulary and randomly initialize the newly added tokens in the embedding matrix of the IndicTrans2 DA models to incorporate representation for these languages. The expanded models serve as the base for fine-tuning.

For our second contrastive system using MEMT, we train 5-gram language models using KenLM (Heafield, 2011) and use default settings for system combination. Instead of taking only the best beam search output of each system being combined, we take the top 2 best translations in the beam, which simulates a combination of 4 systems.

For local evaluation, we use BLEU score (Papineni et al., 2002) measured using sacrebleu (Post, 2018), however, organizers additionally report chrF2 (Popović, 2017), RIBES (Isozaki et al., 2010) and TER (Snover et al., 2006). Human evaluation is not performed, but the organizers release COMET (Rei et al., 2022) scores as an approximation. For test set decoding, we identify optimal decoding hyperparameters (beam size and length penalty) by grid searching on the development set and list said hyperparameters in Table 4 for our primary and first contrastive system.

[4] https://github.com/facebookresearch/fairseq

5 Results

In this section, we describe the results we obtained on the test sets.

5.1 Primary

Main result. Table 5 shows the results of our primary many-to-many system. For the Indic-En direction, Manipuri and Mizo to English exhibit reasonably high translation quality, at BLEU/chrF2 scores of 39.40/64.70 and 32.47/51.33 respectively. Assamese to English translation is the next best at 27.02/50.71. However, Khasi to English has the lowest translation quality among all. A critical observation is that there is no particular correlation between the sizes of the corpora and the MT quality. For example, Manipuri-English has the smallest parallel corpus (21,687 lines) and the second-smallest monolingual corpus (2.2M lines) but still exhibits the best translation quality for Manipuri to English. This could mean that the evaluation set is either easier for this pair or that it is easier to translate the pair compared to others.

For the reverse direction, once again Mizo and Manipuri exhibit the best translation quality, followed by Khasi and Assamese. Despite Assamese having more than 8 million monolingual sentences that were used for backtranslation, its translation quality is at 17.03 and 45.31 (BLEU and chrF2) which is not particularly high. The same decorrela-

Pair	BLEU	chrF2	RIBES	TER	COMET
Primary					
as-en	27.02	50.71	0.71	62.46	0.76
mz-en	32.47	51.33	0.69	60.56	0.67
kha-en	17.80	39.22	0.66	74.10	0.60
mni-en	39.40	64.70	0.77	51.27	0.79
en-as	17.03	45.31	0.58	76.57	0.78
en-mz	33.18	56.73	0.73	55.68	0.70
en-kha	19.95	43.30	0.68	66.47	0.67
en-mni	27.36	61.60	0.74	58.28	0.76
Contrastive #1					
as-en	**37.28**	59.97	0.72	58.81	0.81
mz-en	28.47	47.93	0.61	67.54	0.69
kha-en	**20.06**	40.33	0.58	78.44	0.60
mni-en	**46.06**	69.96	0.80	47.44	0.83
en-as	18.09	51.98	0.57	73.41	0.82
en-mz	26.47	50.60	0.66	65.97	0.69
en-kha	20.77	43.82	0.65	69.51	0.68
en-mni	24.17	62.95	0.70	62.85	0.76
Contrastive #2					
as-en	36.97	59.82	0.72	58.53	0.81
mz-en	**33.30**	52.74	0.70	60.87	0.68
kha-en	20.02	39.82	0.59	77.50	0.59
mni-en	43.35	69.27	0.80	47.43	0.82
en-as	**21.07**	51.71	0.58	73.03	0.81
en-mz	**33.64**	56.88	0.72	57.71	0.71
en-kha	**21.05**	46.06	0.65	73.80	0.68
en-mni	**27.40**	61.55	0.74	58.16	0.76

Table 5: Our primary and contrastive system results for Indic-En and En-Indic translation on the test set. These scores for all the metrics are directly reported as provided by organizers.

Pair	WMT PC	Stage	
		Intermediate	Final
as-en	17.63	24.06	**26.11**
mz-en	22.36	25.98	**28.34**
kha-en	11.03	13.22	**14.68**
mni-en	31.70	36.73	**40.43**
en-as	13.23	16.62	**17.51**
en-mz	21.54	24.25	**26.12**
en-kha	14.72	15.99	**17.60**
en-mni	20.35	23.72	**24.62**

Table 6: Greedy search BLEU scores on the development set for Indic-En and En-Indic direction for the various models we trained in the process of getting to our final model. The “WMT PC" model uses only the parallel corpus for training. The “Intermediate" model is trained using the joint text infilling and MT objective and the “Final" model is trained with the backtranslated and organizers’ parallel data. All models are many-to-many. Please note that we use the IndicNLP tokenizer (Kunchukuttan, 2020) instead of standard tokenizer provided in sacrebleu (Post, 2018) for computing scores locally.

tion between corpora sizes and translation quality that existed for translation into English holds for the reverse direction. In addition, we report the BLEU scores for NLLB 54B MoE model on test set in Table 8.

Ablations. Although we report test set results only using the final system, we also report the BLEU scores on the organizer’s official dev set of the intermediate and final models in Table 6. Additionally, we report the results of a model that is trained only using the organizers’ official parallel corpora. It is clear that the intermediate model using joint denoising and MT training leads to a vast improvement in translation quality, indicating that the monolingual corpus brings substantial benefits. This is especially the case for Indic-En direction since we use around 20M monolingual English sentences. We observe that the En-Indic direction also has some performance gains (around 3 BLEU) but not as much as compared to the gains in the Indic-En direction (around 6 BLEU). This implies that the scale of monolingual data is an important factor, however, we are limited by the scale of monolingual data available for the Indic languages.

Furthermore, the final model, which uses backtranslated data from the intermediate model further shows improvements of approximately 4 BLEU. This indicates that in low-resource settings similar to ours, leveraging monolingual corpora first via denoising followed by backtranslation leads to the best models. Iterative backtranslation (Hoang et al., 2018) would be the ideal next step, but we chose to not pursue it because of compute constraints.

5.2 Contrastive

Contrastive #1: Main Result. Table 5 shows the results of our contrastive #1 system. We observe superior performance for the languages that are already covered in the off-the-shelf IndicTrans2 models (Assamese, Manipuri (Bengali)) as compared to the primary system. For Indic-En direction, Assamese and Manipuri to English exhibit reasonably high translation quality, achieving BLEU scores of 37.28 and 46.06 respectively. Furthermore, we also find mixed results between both systems for newly introduced languages such as Mizo and Khasi. For Khasi, the contrastive #1 system outperforms the primary system on both directions, whereas for

Model variants	FLORES-200 (18 lang)		WMT (all langs)		WMT (new langs)	
	En-Indic	Indic-En	En-Indic	Indic-En	En-Indic	Indic-En
IT2-DA	19.03	37.25	-	-	-	-
A1	3.85	35.81	24.70	32.50	23.20	24.40
A2	19.46	37.62	20.90	24.60	18.95	13.00
A3	18.68	38.07	25.80	32.30	24.50	24.20

Table 7: BLEU scores of different ablations described in Section 3.2 explored under contrastive #1 system on FLORES-200 devtest set (covers 18 languages) and WMT dev set (4 languages). Please note that we use the IndicNLP tokenizer (Kunchukuttan, 2020) instead of standard tokenizer provided in sacrebleu (Post, 2018) for computing scores locally.

	Primary		Contrastive #1		Contrastive #2		NLLB 54B MoE	
	en-xx	xx-en	en-xx	xx-en	en-xx	xx-en	en-xx	xx-en
as	17.03	27.02	18.09	**37.28**	**21.07**	36.97	19.60	26.8
mz	33.18	32.47	26.47	28.47	**33.64**	33.30	27.50	**38.50**
mni	27.36	39.40	24.17	**46.06**	**27.40**	43.35	14.90	31.50

Table 8: Comparison of BLEU scores of all our systems - Primary, Contrastive #1, Contrastive #2 with massively multilingual NLLB 54B MoE model (Costa-jussà et al., 2022).

Mizo, the primary system outperforms the contrastive #1 system across both directions.

Contrastive #1: Ablations. In order to identify the optimal configuration for training the Contrastive #1 system, a series of ablations were conducted, involving a comparative analysis of three distinct approaches for fine-tuning the IndicTrans2 model (Gala et al., 2023), as detailed in Section 3.2. The baseline approach, denoted as A1, focuses solely on optimizing performance across 4 languages under consideration for WMT languages. However, this approach exhibits catastrophic forgetting on the existing supported languages. This is evident in the significant drop in average BLEU scores on the FLORES-200 test set (Goyal et al., 2022; Costa-jussà et al., 2022). Specifically, when fine-tuning IndicTrans2 DA model (Gala et al., 2023) to obtain A1 for the En-Indic language direction, the average BLEU score significantly dropped from 19.03 to 3.85. However, for the Indic-En direction, the drop is relatively modest, shifting from 37.25 to 35.81, although this drop can be made even lower.

To prevent catastrophic forgetting on existing supported languages, an alternative approach, labeled as A2, was experimented. This approach involves a joint fine-tuning on a combined set involving all the existing supported languages along with the newly introduced ones. Notably, this approach averts the issue of catastrophic forgetting. On the FLORES-200 benchmark (Goyal et al., 2022; Costa-jussà et al., 2022), the models resulting from this joint fine-tuning slightly surpass the performance of the IndicTrans2 DA models in both translation directions, showing an improvement of approximately 0.4 points. However, despite this improvement, the performance on the newly added languages such as Khasi and Mizo is suboptimal, significantly trailing behind the scores obtained using the A1 approach. We observe respective drops of 3.8 and 7.9 points in the En-Indic and Indic-En directions over A1.

Although approach A2 successfully resolved the issue of catastrophic forgetting, it did not fully meet our objective of optimizing for the newly introduced languages. As a result, we explored approach A3 which involves a two-stage fine-tuning procedure, wherein A1 is initially employed, followed by A2. As previously discussed, A1 resulted in a sharp decline in performance across the existing languages, but optimized to the newly introduced languages. However, we observe that this performance drop can be rectified by introducing an additional stage of fine-tuning involving a combined set of all languages, as seen in approach A2. A3 results in a fair retention of performance in terms of BLEU scores across the existing languages for both translation directions, Indic-En (with scores of 37.24 for IndicTrans2 DA and 36.68

for A3) and En-Indic (with scores of 19.03 for IndicTrans2 DA and 18.68 for A3). Moreover, on the newly introduced languages, models trained using the A3 approach demonstrate an improvement of nearly 8 points in the Indic-En direction and 4.9 points in the En-Indic direction on average, when compared to A2, as observed on the WMT 2023 IndicMT dev set. Notably, A3 achieves performance on par with A1 (optimized for four languages) in the Indic-En direction and even outperforms A1 by a margin of +1.1 in the En-Indic direction. Therefore, A3 achieves both the outcomes: performance retention on existing languages as well as optimization in performance for newer languages.

Contrastive #2: Main Result. Having obtained the best primary and contrastive systems, we combine them via MEMT. Table 5 contains the result of the system combination on the test set. For Indic-En direction, only Mizo to English benefits from system combination, where the best BLEU score improves from 32.47 to 33.30. For the En-Indic direction, we see improvements for all directions. Most notable is the improvement for English to Assamese, whose best BLEU score improves from 18.09 to 21.07. For other directions, the improvements are relatively smaller. One important observation is that when the performance gaps between the primary and contrastive #1 system is larger, the gains are smaller or are negative. Overall, it is important to note that such word level system combination still works despite the idea being over a decade old, however, the use of n-gram based LMs might be a limitation and replacing said LMs with neural LLMs might bring large benefits. We leave this for future work.

5.3 Lessons Learned

- In low-resource settings, leverage monolingual data first via denoising and then via backtranslation.
- A two-stage fine-tuning approach (introducing new languages first, followed by a combination of new and existing languages) is an effective approach when considering extending a pre-trained translation model to newer languages without catastrophic forgetting.
- System combination is still effective despite working at a word level.

6 Conclusion

In this paper, we have described our systems submitted to the WMT 2023 Indic translation task. We leveraged ideas ranging from joint denoising and MT training, backtranslation, fine-tuning pretrained models, and system combination. We reported our results, which show the benefits of the various ideas we explored. Finally, we recommend best practices.

7 Limitations

We identify the following limitations of our submissions:

- We did not perform ensembling or checkpoint averaging, which could boost our results by another 1-2 BLEU.
- Iterative backtranslation (Hoang et al., 2018) was not adopted due to compute constraints and can potentially boost quality even further.
- Although we reached the monolingual corpora limit for the Indic languages of focus, we could have used much larger English monolingual corpora but opted not to, once again, due to compute constraints. This would also require us to increase model sizes which was also not feasible.
- We have not leveraged any LLMs for our experiments, mainly because we are not sure if they have been trained on any of the test data, a common concern in recent times.
- MEMT is an old idea and does not use any neural language models, especially LLMs, which could enhance its performance.

Acknowledgement

We graciously thank Varun Gumma (SCAI Fellow, Microsoft Research) for his kind review of this draft, which helped us improve its readability and quality.

References

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *CoRR*, abs/1907.05019.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5).

Raj Dabre, Diptesh Kanojia, Chinmay Sawant, and Eiichiro Sumita. 2023. YANMTT: Yet another neural machine translation toolkit. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 257–263, Toronto, Canada. Association for Computational Linguistics.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91, Edinburgh, Scotland. Association for Computational Linguistics.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Jay Gala, Pranjal A. Chitale, Raghavan AK, Sumanth Doddapaneni, Varun Gumma, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv: 2305.16307*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Kenneth Heafield and Alon Lavie. 2010. CMU multi-engine machine translation for WMT 2010. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 301–306, Uppsala, Sweden. Association for Computational Linguistics.

Kenneth Heafield and Alon Lavie. 2011. CMU system combination in WMT 2011. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 145–151, Edinburgh, Scotland. Association for Computational Linguistics.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952, Cambridge, MA. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Samta Kamboj, Sunil Kumar Sahu, and Neha Sengupta. 2022. DENTRA: Denoising and translation pre-training for multilingual machine translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1057–1067, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Anoop Kunchukuttan. 2020. The IndicNLP Library. https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzman. 2023. Small data, big impact: Leveraging minimal data for effective machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, Toronto, Canada. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Sandeep Kumar Dash, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, and Pankaj Kundan Dadure. 2023. Findings of the wmt 2023 shared task on low-resource indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

Machine Translation Advancements for Low-Resource Indian Languages in WMT23: CFILT-IITB's effort for bridging the Gap

Meet Doshi, Pranav Gaikwad, Sourabh Deoghare, Pushpak Bhattacharyya
Computation for Indian Language Technology Lab
Indian Institute of Technology, Bombay.
{meetdoshi, pranavgaikwad, sourabhdeoghare, pb}@cse.iitb.ac.in

Abstract

Machine translation for low-resource Indian languages has long been a challenge due to the scarcity of high-quality parallel corpora, demanding the development of effective translation models. The WMT23 Low-Resource Indic Language Translation task encourages us to utilize creative techniques to address this issue and enhance the performance of machine translation systems for these languages. We focused on the translation of two low-resource Indic languages: Assamese and Manipuri, enabling bidirectional translation between English and these languages. This paper presents CFILT-IITB's submission to WMT23, highlighting our exploration of transfer learning-based methodologies. Our experiments produced notable results of **47.54 BLEU** on MNI→EN, **18.15 BLEU** on EN→ASM and **35.24 BLEU** on ASM→EN, **26.36 BLEU** on EN→MNI test sets. These results not only demonstrate the effectiveness of transfer learning-based techniques but also contribute to advancing machine translation capabilities for low-resource Indian languages, addressing a critical need in bridging language barriers and facilitating cross-cultural communication.

1 Introduction

In the realm of machine translation, the WMT23 IndicMT shared task emerges as an arena where the boundaries of translation technology are stretched to their limits. Our efforts revolve around the translation of the 'En-X' pair in both directions, where 'En' signifies English and 'X' encompasses Assamese, a member of the Indo-Aryan language family, and Manipuri, a representative of the Tibeto-Burman family. As the task focused on English to and from low-resource Indian languages, we were provided with a small parallel corpus for each 'En-X' pair. Furthermore, participants had access to a substantial amount of monolingual data for Assamese and Manipuri, creating an ideal setting for trying out new and creative approaches.

In the realm of Machine Translation, the Neural Machine Translation paradigm has emerged as a dominant force, as evidenced by seminal works such as (Bahdanau et al., 2014) and the comprehensive survey by (Dabre et al., 2020). However, Neural Machine Translation models are notoriously data-hungry, leading to performance degradation when confronted with low-resource languages, as highlighted by (Dewangan et al., 2021). To tackle this challenge, we turn to the promising technique of transfer learning, a well-established approach in machine learning where knowledge gained from one task is leveraged to enhance performance in a related task. In our pursuit of improving translation capabilities for low-resource languages, we harness the multilingual IndicTrans2 model, as introduced by (AI4Bharat et al., 2023). Our methodology involves fine-tuning this model using the 'En-X' parallel data provided for the task. By adopting this approach, we aim to capitalize on the acquired knowledge during training to significantly bolster the performance of the model in the specific translation task at hand.

IndicTrans2 is rooted in the transformer-based encoder-decoder architecture pioneered by (Vaswani et al., 2017). It was trained on the extensive Bharat Parallel Corpus Collection (BPCC), a publicly accessible repository encompassing both pre-existing and freshly curated data for all 22 scheduled Indian languages, this model boasts a comprehensive understanding of the linguistic diversity within the Indian subcontinent. To enhance its linguistic prowess, IndicTrans2 has undergone auxiliary training utilizing the rich resource of back-translated monolingual data. The model was then trained on human-annotated data to achieve further improvements. We used this model and fine-tuned it on the training data provided by WMT23.

The fine-tuned IndicTrans2 achieves good scores; hence we are using it for our final submission. We hypothesize that its stellar performance

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 950–953
December 6–7, 2023.

can be attributed to the amalgamation of language knowledge acquired during its initial training, coupled with the domain-specific expertise gleaned from the fine-tuning process, facilitated by the training data made available through WMT23.

2 Data

We use the IndicTrans2 model and fine-tune it on the WMT23. The original IndicTrans2 was trained on the Bharat Parallel Corpus Collection (BPCC) corpus. They have used FLORES-200 as their validation set for Assamese and extended FLORES-200 (Team et al., 2022) for Manipuri. For auxiliary training which includes back-translated monolingual sentences, they have used IndicCorp v2 (Kakwani et al., 2020) and one side of NLLB data as monolingual corpus. They have used standard test sets like FLORES-200, but they have also created a new benchmark called the IN22 test set which is an n-way parallel corpus for all 22 Indian scheduled languages.

We have fine-tuned the model using the WMT23 parallel corpus. The 'English-Assamese' pair has 50K parallel sentences, and the 'English-Manipuri' pair has around 21.6K sentences. The validation set consisted of the WMT23 validation set. The size of the validation set for the 'English-Assamese' is 2K sentences; for the 'English-Manipuri' pair, it was 1k sentences. The test set for both pairs was the WMT23 test set.

3 System Overview

In the pursuit of enhancing machine translation for low-resource languages, various approaches have emerged, such as Phrase-Pair injection and Back-translation, aimed at enhancing performance. Our system, on the other hand, takes a distinct path and relies on the knowledge gained from the multilingual training of IndicTrans2 and applies it to different low-resource languages.

Phrase-Pair Injection (PTI) (Sen et al., 2021), (Dewangan et al., 2021) and (Banerjee et al., 2021) utilized a technique to combine both Statistical Machine Translation (SMT) and Neural Machine Translation (NMT). The utilization of the phrase table during training is pivotal in Statistical Machine Translation (SMT) as it probabilistically maps phrases from the source to the target language. By incorporating these phrase mappings from the table into the existing parallel corpora, the training dataset for the Neural Machine Translation (NMT) model is significantly enriched. Consequently, this enrichment empowers the NMT model to excel in its translation performance.

Back-translation Back-translation (Sennrich et al., 2016; Conneau et al., 2020) is a technique that is used to improve the performance of low-resource translation systems using monolingual data. In this technique, a reverse model is employed to generate a parallel corpus from a monolingual corpus. This is a clever way to use the monolingual corpus to improve the translation performance of the NMT models. We do not include Back-translated sentences for training since we could not see any significant performance improvement.

Transfer Learning Transfer learning is a machine learning technique where a model trained on one task is adapted for a second related task. Instead of starting the training of a new model from scratch, transfer learning leverages the knowledge learned from the first task to improve learning on the second task. We have used IndicTrans2 (AI4Bharat et al., 2023), a powerful model that performs well for English-to-Indic and Indic-to-English translation for 22 scheduled Indian languages. This knowledge can be used to translate other Indian languages to and from English. Our approach entailed the fine-tuning of this model, leveraging the parallel corpus provided by the WMT23 for the IndicMT task. This fine-tuning process equipped the model with the expertise required to proficiently translate Assamese and Manipuri to and from English, ultimately yielding the most outstanding results. We do not inject phrase pairs since for such a low resource setting, it is difficult to see performance improvements even with phrase pair injections due to the inability of NMT models to capture the low resource language.

4 Experiments

4.1 Settings

All the experiments are conducted using two NVIDIA A100 GPUs each having 80GB of memory. Our models apply Adam (Kingma and Ba, 2015) as optimizer to update the parameters with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. We employ a warm-up learning rate of 10^{-7} for 2000 update steps and a learning rate of $3 * 10^{-5}$. For normalization, we use a dropout value of 0.2 and normalize the proba-

Models	ASM→EN		EN→ASM		MNI→EN		EN→MNI	
	BLEU	**ChrF2**	**BLEU**	**ChrF2**	**BLEU**	**ChrF2**	**BLEU**	**ChrF2**
Baseline-1 (val)	2.32	-	1.64	-	3.12	-	2.67	-
IndicTrans2 (val)	25.60	47.20	14.70	41.40	33.40	58.50	11.90	43.50
FT IndicTrans2 (val)	34.60	52.40	**24.00**	46.00	47.00	67.30	**34.10**	62.20
FT IndicTrans2 (test)	**35.24**	**57.73**	18.15	**50.16**	**47.54**	**70.41**	26.36	**63.48**

Table 1: Comparison of results of Fined-tuned IndicTrans2 (AI4Bharat et al., 2023) on the test and val set. We compare val and test set results because we see that the EN-Indic model has overfitted for both languages and therefore we see a decrease in BLEU for EN-Indic models. We recommend readers to decrease the number of updates for better scores when the source is English.

bilities using smoothed label cross-entropy. We use GeLU activations (Hendrycks and Gimpel, 2016) for better learning. We train separate models for each language pair to avoid data imbalance and learn better low-resource representations.

We use the scareBLEU library[1] to calculate our BLEU (Papineni et al., 2002) and ChrF (Popović, 2015) scores with a word order of 2. We choose the checkpoint with the highest validation BLEU score.

4.2 Results

Table 1 shows that the highest translation quality achieved is via the use of large monolingual and parallel corpora. Since IndicTrans2 is trained in many Indian languages, it enhances the translation quality via the power of multilingualism. With only some minor tuning of the model over the training and validation set, IndicTrans2 achieves remarkable performance on Indic-En translations. Our baseline-1 system is a WMT-14 En-De fairseq model trained that utilizes only the parallel data and shows substandard BLEU scores over all the language pairs. With our experiments, we see that with even the monolingual corpora and back translation, the translation models only see minor improvements. We realized the power of multilingualism and switched to pre-trained models which have been trained on a substantial amount of data like IndicTrans2 (AI4Bharat et al., 2023) and NLLB (Team et al., 2022). We analyze their vocabulary and merge it with a new vocabulary learned over the monolingual corpora provided in the task. Even for languages that are not seen by the model like Mizo and Khasi in the *latin* script, the IndicTrans2 model with its pre-trained English vocabulary gives a BLEU score of an average of 7.2 on the val set over these language pairs. We see that when we fine-tune the pre-trained model, we see large gains over both the *val* and the *test* set. Finally, after many experiments, we submit a fine-tuned version of a very powerful multilingual model for the shared task.

[1] https://github.com/mjpost/sacrebleu/blob/master/sacrebleu/metrics/bleu.py

5 Conclusion

In this paper, we present how CFILT-IITB utilized the power of multilingual models for the WMT23 IndicMT Low-Resource Machine Translation of Indian Languages Shared Task. Since, the data for low-resource languages is scarce, utilizing pre-trained multilingual translation models is very crucial. But to have reasonable to good performance over these models, it is helpful to have a model that is trained on similar languages. For this task, Indian languages like *Assamese* and *Manipuri* share similar structure and vocabulary with many Indian languages like *Bengali* which can be considered a high resource language for India. Training models over similar language does boost performance although to cover a wide variety of low-resource languages, one must face the curse of multilingualism. Our most proficient system attains an average BLEU score of 41.39 for Indic-English translation and 22.25 for English-Indic language pairs, specifically Assamese and Manipuri.

Limitations

Limitations of such powerful multilingual models are data extraction, enormous computing, and good data filtration techniques. Overcoming these obstacles is an open research problem.

References

AI4Bharat, Jay Gala, Pranjal A. Chitale, Raghavan AK, Sumanth Doddapaneni, Varun Gumma, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M.

Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Aakash Banerjee, Aditya Jain, Shivam Mhaskar, Sourabh Dattatray Deoghare, Aman Sehgal, and Pushpak Bhattacharyya. 2021. Neural machine translation in low-resource setting: a case study in english-marathi pair. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 35–47.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Computing Surveys (CSUR)*, 53(5):1–38.

Shubham Dewangan, Shreya Alva, Nitish Joshi, and Pushpak Bhattacharyya. 2021. Experience of neural machine translation between indian languages. *Machine Translation*, 35(1):71–99.

Dan Hendrycks and Kevin Gimpel. 2016. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, abs/1606.08415.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Sukanta Sen, Mohammed Hasanuzzaman, Asif Ekbal, Pushpak Bhattacharyya, and Andy Way. 2021. Neural machine translation of low-resource languages using smt phrase pair injection. *Natural Language Engineering*, 27(3):271–292.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Low-Resource Machine Translation Systems for Indic Languages

Ivana Kvapilíková and **Ondřej Bojar**
Institute of Formal and Applied Linguistics,
Faculty of Mathematics and Physics, Charles University,
Prague, Czech Republic,
kvapilikova@ufal.mff.cuni.cz, bojar@ufal.mff.cuni.cz

Abstract

We present the submission of the CUNI team to the WMT23 shared task in translation between English and Assamese, Khasi, Mizo, and Manipuri. All our systems were pretrained on the task of multilingual masked language modelling and denoising auto-encoding. Our primary systems for translation into English were further pretrained for multilingual MT in all four language directions and fine-tuned on the limited parallel data available for each language pair separately. We used online back-translation for data augmentation. The same systems were submitted as contrastive for translation out of English where the multilingual MT pretraining step seemed to harm the translation performance. Other contrastive systems used additional pseudo-parallel data mined from monolingual corpora.

1 Introduction

We present our submission to the Indic MT shared task of the WMT23 workshop. We trained constrained systems in all evaluated language directions: English-Assamese (en-as), English-Manipuri (en-mni), English-Mizo (en-mz) and English-Khasi (en-kha).

A majority of languages in the world have a very limited amounts of translation resources to be used for training machine translation (MT) systems. Unsupervised learning techniques have been proposed to leverage monolingual texts in MT training, either in the pretraining phase (Liu et al., 2020; Conneau and Lample, 2019) or during fine-tuning by means of back-translation (Sennrich et al., 2016). This shared task is proposed as a realistic scenario where for each Indic language, the participants have access to several thousand parallel sentences paired with English and up to 2.6M additional unaligned sentences in each language. The texts are mixed from the religious domain and the general domain. In addition to the provided data, participants were allowed to use any monolingual texts and any pretrained models trained on monolingual texts.

In our other research, we focus primarily on unsupervised MT and we participated in this shared task to measure the impact of adding at least a small number of parallel sentences into the training. Therefore, we also evaluated our fully unsupervised systems in the conditions of Indic MT, where the languages are linguistically very different from English and some also have a different script.

A major obstacle, especially for our unsupervised models, is the domain mismatch in our training data. While monolingual English data we used come from NewsCrawl, the Indic training data includes texts from the religious domain. The issue is especially pronounced when we struggle with finding equivalent sentences in the monolingual corpora section 2.5, but it is problematic for the entire unsupervised training as the domain mismatch interferes with the underlying assumption of isomorphism of embedding spaces.

In this paper, we first introduce our training methodology (section 2), describe the data (section 3.1) and comment on the results (sections 4 and 5)

2 Methodology

2.1 Model Architecture

The architecture of all our NMT models is a 6-layer Transformer with 6 attention heads, GELU activations, and 0.1 dropout. In addition to token embeddings and trained positional embeddings, the model features language embeddings to pass the information which language direction is being used. Both input token embeddings and the final softmax layer have tied weights.

2.2 Pretraining on Monolingual Texts

We pretrain a Transformer encoder on the task of masked language modelling (MLM) on all available corpora in all languages. The details of

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 954–958

	as	kha	mni	mz	en
train (mono)	2.6M	183k	2.1M	1.9M	33M
train (para)	50k	24k	22k	50k	-
train (pseudo-para)	81k	95k	150k	66k	-
dev	2k	1k	1k	1.5k	-
test	2k	1k	1k	2k	-

Table 1: The number of sentences in the training, dev and test sets. Monolingual (mono) and parallel (para) data were provided by the organizers, pseudo-parallel data was created as described in Section 2.3.

the MLM task are given in Conneau and Lample (2019). We copy its weights into both the encoder and the decoder of the Transformer model and we continue the pretraining phase by training a multilingual denoising autoencoder (DAE). The noise function applied to the input sentence has the following components: word deletion with probability $p_{del} = 0.1$, word masking with probability $p_{mask} = 0.1$ and word shuffling within the window of length $l_{shuf} = 4$.

All our systems are pretrained on both MLM and DAE.

2.3 Pretraining on Multilingual Parallel Texts

In the second pretraining stage, we train a multilingual neural machine translation model (MNMT) on all available parallel data. In each training step, the model sees a mini-batch of parallel sentences for all language pairs. It uses language embeddings to detect the right translation direction.

2.4 Fine-tuning for Machine Translation

In the fine-tuning stage, we train a bidirectional model for each language pair in a semi-supervised fashion, using a cross-entropy loss on a small authentic parallel corpus. We augment the data with online back-translation (OBT) (Lample et al., 2018; Artetxe et al., 2018) to avoid over-fitting. In every OBT training step, the model is switched into an inference mode to create a mini-batch of training data by translating a portion of monolingual sentences. This operation is performed in both translation directions and the resulting mini-batch (with the synthetic sentences placed on the source side) is directly used for training.

2.5 Data Augmentation with Pseudo-Parallel Texts

We also measure whether we can earn some benefits by incorporating pseudo-parallel (PP) sentences into the MT training. We use the methodology of Kvapilíková et al. (2020) and search for parallel sentences in the training corpora. We search for the nearest neighbors in the multilingual sentence embedding space created by a multilingual sentence encoder. The encoder is the modified XLM-100 (Conneau and Lample, 2019) pretrained model fine-tuned on the MLM task for Assamese, Khasi, Mizo, Manipuri and English. The search metric is the modified cosine similarity $xsim$ (Artetxe and Schwenk, 2019) between the sentence embeddings which is required to be higher than 1:

	en-as	**en-kha**	**en-mni**	**en-mz**
Precision	35.03	9.67	7.92	22.54
Recall	18.55	10.50	5.70	18.00
F1 Score	24.26	10.07	6.63	20.01
Threshold	1.022	1.027	1.022	1.022

Table 2: Precision, Recall and F1 score on the Parallel Sentence Matching Task.

$$\text{xsim}(x, y) = \frac{\cos(x, y)}{\text{avgcos}(x) + \text{avgcos}(y)} > 1 \quad (1)$$

where

$$\text{avgcos}(\cdot) = \sum_{z \in \text{NN}_k(\cdot)} \frac{\cos(\cdot, z)}{2k} \quad (2)$$

where $\text{NN}_k(x)$ is the set of k nearest neighbors of x. We augment the existing authentic parallel corpora with the pseudo-parallel sentence pairs and train on the resulting corpus. The number of retrieved pseudo-parallel sentence pairs is indicated in table 1. The performance of the sentence encoder at the task of parallel corpus mining for the languages in question was measured by an auxiliary task where it was asked to find 1-2k parallel sentences (dev set) among 200k monolingual sentences from the train set in both languages. The results are summarized in section 2.4 where we see that the precision of correctly matched parallel sentences for Khasi and Manipuri is very low.

	en-as	en-kha	en-mni	en-mz	as-en	kha-en	mni-en	mz-en
MT+OBT	**14.1**	**16.6**	29.5	31.2	17.6	12.8	33.9	28.3
MNMT+MT+OBT	13.9	16.4	**29.9**	**31.5**	**20.7**	**13.8**	**36.1**	**29.5**
PP+MT+OBT	13.3	15.9	29.8	30.8	16.8	12.1	30.2	28.7
OBT (unsup)	0.2	-	-	0.8	0.3	-	-	1.3
PP+OBT (unsup)	2.9	-	-	6.1	3.1	-	-	5.5

Table 3: BLEU score of our MT systems on the WMT23 test set.

In our experiments we evaluate the impact of MNMT and PP pretraining on the final translation quality.

3 Experiments

We train several models for each language pair. All models are pre-trained as described in 2.2. For our shared task submission, we train three kinds of semi-supervised models using all available parallel data:

- MNMT+MT+OBT models were trained for multilingual MT and fine-tuned for each language pair separately on a combination of authentic parallel data and synthetic parallel data created by OBT;
- MT+OBT models skip the multilingual MT pre-training step;
- PP+MT+OBT models are trained on pseudo-paralle data in addition to authentic and synthetic data. The pseudo-parallel corpus is removed after 5 epochs of training.

We compare the results of the semi-supervised models to unsupervised models trained without the authentic parallel data to measure the effect of limited amounts of parallel data. We experiment with gradually adding parallel sentences into the training and evaluate the performance of a model trained on 1k, 2k, 5k, 10k and 25k parallel sentences.

3.1 Data

In addition to the data provided by the organizers (Pal et al., 2023), we used 33M English sentences from NewsCrawl2022. The summary of the data is in table 1. We trained a BPE model on the concatenation of all Indic corpora and a downsampled Englih corpus. The BPE vocabulary size is 52k. During pre-processing, we first tokenized the texts using the Moses tokenizer which created a problem with the Assamese script as it decomposed several compound Unicode characters which had impact on the segmentation of texts with Assamese script (as, mni). The decomposed accents form a separate BPE unit which lead to a high segmentation of the Assamese and Manipuri texts. During post-processing we managed to compose the segmented text by running a special substitution on top of the standard detokenization. The unnecessary step of Moses tokenization likely cost us some final translation performance due to the sub-optimal BPE segmentation.

3.2 Training Details

We use the XLM[1] toolkit for training. For language model pretraining, we use mini-batches of 64 text streams (256 tokens per stream) per GPU and Adam (Kingma and Ba, 2015) optimization with `lr=0.0001`. For denoising and MT fine-tuning, we use mini-batches of 3,400 tokens per GPU and Adam optimization with a linear warm-up `(beta1=0.9,beta2=0.98,lr=0.0001)`. The models are trained on 8 GPUs.

4 Shared Task Results

For our shared task submission, we compared the performance of our experiments on the dev set and concluded that the fine-tuned multilingual NMT system (MNMT+MT+OBT) performs better than individual systems (MT+OBT) when translating into English and on par with individual systems when translating from English. Therefore, for our PRIMARY submission, we submitted the output of the multilingual model when translating into English and the output of the individual models when translating from English. The opposite results were submitted as CONTRASTIVE-1. The PP+MT+OBT systems were submitted as CONTRASTIVE-2. The final test set results are summarized in table 3.

The winning system for all language directions was a system called TRANSSION-MT which outperformed other systems with almost double the

[1] https://github.com/facebookresearch/XLM

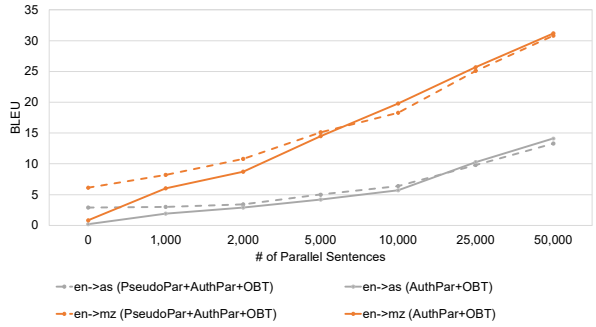


Figure 1: Relationship between the translation quality and the number of parallel sentences used for training.

BLEU score of the second best candidate (Pal et al., 2023). In general, our systems performed relatively better in translation from English which suggests that the translation to English may have been harmed by the bidirectional nature of our systems. Our en→mni system ranked second after TRANSSION-MT out of 14 participants. Our en→mz system ranked fourth out of 11 participants. The the remaining systems finished on the 5th-7th places.

5 Discussion

Asides from the shared task submission, we were interested in the following phenomena which we measured in our experiments:

- The gap between unsupervised and semi-supervised translation systems;
- The impact of pseudo-parallel data augmentation on the final translation quality;
- The development of translation quality in relation to the number of parallel sentences used during training.

Outside of the scope of the shared task, we trained unsupervised MT systems for Mizo and Assamese. For each of the two language pairs, we trained two systems, with and without pseudo-parallel sentences. table 3 shows that the unsupervised systems reach between 3 and 6 BLEU which is not a sufficient quality for practical use of the systems. The poor unsupervised results are most likely the consequence of the domain mismatch between English and Indic data as well as a mismatch between the English train set and the test sets. Our conclusions support the claims of other researchers (Marchisio et al., 2020; Vulić et al., 2019) that unsupervised MT models often fail in truly low-resource scenarios where it is not possible to obtain enough clean and domain-balanced monolingual training data and the underlying assumption of language isomorphism is challenged.

Data augmentation with pseudo-parallel sentences has zero or even a negative impact on the performance of our semi-supervised systems. For the unsupervised systems, on the other hand, it increases BLEU score by up to 5.3 BLEU points. We trained several other systems, gradually adding more parallel sentences, to measure the threshold where pseudo-parallel sentences stop helping. fig. 1 illustrates the relationship between translation quality and reveals that when we have more than 10k parallel-sentences, the unsupervised data augmentation techniques of adding pseudo-parallel sentence pairs is not beneficial anymore.

6 Conclusion

We trained several MT systems for translation between English and four Indic languages. The most promising outcomes were achieved by initially pre-training a multilingual NMT system, followed by fine-tuning using bilingual parallel data along with online back-translation. Our systems ranked between the 2th and 7th place among 10-14 participating teams, depending on the language pair and translation direction. Our systems performed relatively better at translation out of English.

Data augmentation with pseudo-parallel data does not bring any further benefits in the context of the shared task. Our experiments show that their positive effect disappears when we have access to more than 10k authentic parallel sentences.

We compared the results to completely unsupervised systems and we conclude that the domain mismatch between our English and Indic training data and the linguistic dissimilarity of the languages do not allow the unsupervised MT systems to learn to translate without seeing parallel sentences. Incorporating pseudo-parallel sentences into the training helps, but the translation quality remains low.

Acknowledgements

This research was partially supported by the grant 19-26934X (NEUREM3) of the Czech Science Foundation and by the SVV project number 260 698 of the Charles University.

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*.

Mikel Artetxe and Holger Schwenk. 2019. Margin-based parallel corpus mining with multilingual sentence embeddings. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32*, pages 7059–7069. Curran Associates, Inc.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations*.

Ivana Kvapilíková, Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Ondřej Bojar. 2020. Unsupervised multilingual sentence embeddings for parallel corpus mining. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 255–262, Online. Association for Computational Linguistics.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the 6th International Conference on Learning Representations*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work? In *Proceedings of the Fifth Conference on Machine Translation*, pages 571–583, Online. Association for Computational Linguistics.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Sandeep Kumar Dash, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, and Pankaj Kundan Dadure. 2023. Findings of the wmt 2023 shared task on low-resource indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4407–4418, Hong Kong, China. Association for Computational Linguistics.

MUNI-NLP Systems for Low-resource Indic Machine Translation

Edoardo Signoroni and **Pavel Rychlý**
Faculty of Informatics
Masaryk University
e.signoroni@mail.muni.cz, pary@fi.muni.cz

Abstract

The WMT 2023 Shared Task on Low-Resource Indic Language Translation featured to and from Assamese, Khasi, Manipuri, Mizo on one side and English on the other. We submitted systems supervised neural machine translation systems for each pair and direction and experimented with different configurations and settings for both preprocessing and training. Even if most of them did not reach competitive performance, our experiments uncovered some interesting points for further investigation, namely the relation between dataset and model size, and the impact of the training framework. Moreover, the results of some of our preliminary experiments on the use of word embeddings initialization, backtranslation, and model depth were in contrast with previous work. The final results also show some disagreement in the automated metrics employed in the evaluation.

1 Introduction

This paper describes our systems to the WMT 2023 Shared Task on Low-Resource Indic Language Translation. The task featured four low-resource languages indigenous to the northeastern regions of the Indian subcontinent. The translation was to be done to and from Assamese (Indo-Aryan), Khasi (Austroasiatic), Manipuri, Mizo (Sino-Tibetan) on one side and English on the other. We submitted supervised neural machine translation systems for each pair and direction and experimented with different configurations and settings for both preprocessing and training. We did not use large pretrained models, but trained transformers (Vaswani et al., 2017) of different size and with different parameters for each direction, both on bilingual and multilingual data. Even if most of our final systems did not reach a satisfactory or competitive performance, settling for the middle to low part of the scoreboard, we argue that our experiments brought up some interesting points that call for a deeper investigation. Chiefly, these are the relation between dataset and model size, and the impact of the training framework. Moreover, the final results seem to confirm recent research on the reliability of automatic evaluation metrics, with several cases of disagreements in the ranking of the systems, ranging from one to several places in the leaderboard.

2 Datasets

In the following Section, we first briefly present the languages involved, then we give a summary of the datasets, their contents, domains, and structure.

2.1 Languages

Assamese (*Asamiya*) is an Indo-Aryan language mainly spoken by more than 15 million people in the Indian state of Assam, where it is also the official language. Assamese is also one of the 22 official languages recognized by the Republic of India at the federal level. It is influenced by several other regional languages, mostly Tibeto-Burman varieties, and Bengali, another Indo-Aryan language, with which shares the Bengali-Assamese writing script, an abugida system. Assamese serves as a quasi *lingua franca* for the region and functions as one of the source languages for some pidgins and creoles of the area, such as Nefamese and Nagamese. Assamese is an inflected language with eight grammatical cases and a large collection of classifiers. It follows the subject-object-verb order.

Khasi (*Ka Ktien Khasi*) is an Austroasiatic language with 1 million speakers (the Khasi people) in the Indian state of Meghalaya. It is official in some districts of the state, but not in the state as a whole, and it is considered as "vulnerable". It is related with the other languages in the Khasian group native to the Shillong Plateau, and it is surrounded by unrelated languages such as Assamese, Bengali, Manipuri, and others. It is written both in the Latin, as is the case with this task's data, and the Bengali scripts. Khasi is a stress language

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 959–966

Language	ISO-639-3	Family	Script	Num. of Speakers	Official at:	Vitality
Assamese	asm	Indo-Aryan	Bengali	15M	Federal	-
Khasi	kha	Austronesian	Latin	1M	Local	VUL
Mizo	lus	Sino-Tibetan	Latin	1.8M	State	-
Manipuri	mni	Sino-Tibetan	Meitei, Bengali	0.85M	Federal	VUL

Table 1: Summary of the Indic languages involved in the task. For each language, the columns give its ISO code, its language family, the writing system(s) it employs, the number of its speakers and its status, both in terms of official recognition and conservation according to the UNESCO. Khasi and Manipuri are listed as "Vulnerable".

without tones. It has nine grammatical cases and follows the subject-verb-object word order.

Mizo (*Mizo ṭawng*) is a Tibeto-Burman language spoken by around 850 thousand Mizo people, primarily in the Indian state of Mizoram, where it is an official language. It is written in a modified version of the Latin script. Mizo is a tonal language with eight tones, it follows the object-subject-verb order, and it has six grammatical cases.

Manipuri (*Meiteilon*) is a Tibeto-Burman language official in the Manipuri state of India and also at federal level. It is spoken natively by 1.8 million people, the Meitei, both in Manipur and in small communities in the neighboring states. It is considered "vulnerable" by the UNESCO. Manipuri employs a wide array of writing systems, the official ones being the Meitei script and the Bengali script.[1] The Latin script is also used. Manipuri is a tonal language, It follows the subject-object-verb word order.

Table 1 summarizes the main facts about this task's Indic languages.

2.2 Composition

Table 2 gives, for each language pair, the size of the datasets. The parallel datasets made available for this task are small, with the biggest being *asm* and *lus*, at 50k sentence pairs. Of the Indic languages, two are written in Bengali script (*asm* and *mni*), and two in their own variations of the Latin script (*kha* and *lus*). Following the notation of the Flores-200 benchmark dataset (Goyal et al., 2022), we denote the collation of data in Bengali script with *Beng*, and in Latin script with *Latn*.

Monolingual data was released for all Indic languages: *asm*,*lus*, and *mni* have around 2/2.5M sentences each, while *kha* has only 180k. While we did not look at the domains for these data, we sampled the content of the parallel datasets.

[1]This is the writing system used in the task's Manipuri dataset.

Table 3 gives an outline of the contents of each split of each dataset. For *asm*, both the valid and test set differ from the training data. The former is composed mainly by dictionary definitions, while the latter mostly contains religious content. The *kha* dataset is consistent in terms of domains. The *lus* train split has almost exclusively religious content, the validation split contains both religion and instances of single words, and the test split is quite mixed in content. The *mni* data is almost entirely composed by news or otherwise informative text.

3 Methodology

This Section describes our methodology and the baselines we moved from.

3.1 Baselines

For our experiments, we set as our baseline a standard Transformer (Vaswani et al., 2017) with the hyperparameters in Table 4. We wanted to experiment with ways to make the most out of the training data given, and thus we did not use pre-trained models in our work. Almost all models were trained with Fairseq (Ott et al., 2019). The two final submissions to and from Assamese were trained with TorchScale (Ma et al., 2022).

3.2 Preprocessing

The preprocessing for our models was done with SentencePiece (Kudo, 2018), both BPE (Sennrich et al., 2016) and Unigram (Kudo, 2018), and HFT (Signoroni and Rychlý, 2022a). We chose these three segmentation algorithms either for their popularity, as it is the case with BPE and Unigram, or for their stated application, in the case of HFT. For all these algorithms, we set as our baseline parameters a vocabulary size of 2000, with separate dictionaries for source and target language, and a frequency threshold of 100. For other experimental and final runs, we explored different values and settings of segmentation algorithm, vocabulary size and learning, and frequency threshold.

Dataset	Train	Valid	Test	Monolingual	Script
eng-asm	50,000	2,000	2,000	2,624,715	Beng
eng-kha	24,000	1,000	1,000	182,737	Latn
eng-lus	50,000	1,500	2,000	1,909,823	Latn
eng-mni	21,687	1,000	1,000	2,144,897	Beng
eng-Beng	71,687	200	-	-	Beng
eng-Latn	74,000	200	-	-	Latn
eng-all	145,687	400	-	-	Both

Table 2: Size of the dataset for each language pair. Languages are given in ISO-639-3 codes. Train, valid, and test splits are in number of sentence pairs, whereas monolingual data are in number of sentences for the target language. To denote the collation of languages that in the task data are written in Bengali script (asm and mni) and Latin script (kha and lus), we use *Beng* and *Latn*, respectively. *all* denotes the collation of all train splits.

Dataset	Domain		
	Train	Valid	Test
eng-asm	rel,news	misc,news	misc
eng-kha	rel	rel	rel
eng-lus	rel	rel,misc	misc,rel
eng-mni	news	news	news,misc

Table 3: Domains contained in each split of each dataset. Our investigation was conducted on a random sample of each split. *rel(igion)* denotes Bible text and religious news; *news* stands for all non-religious news and information; and *misc* indicates all other miscellaneous domains, e.g. short conversational phrases, dictionary definitions, words.

Parameters	
encoder/decoder layers	6
enc/dec embedding dim	512
enc/dec feed forward dim	2048
enc/dec attention heads	8
optimizer	adam
learning rate	1e-3
warmup updates	4000
dropout	0.3
label smoothing	0.1
max tokens	16384

Table 4: Hyperparameters for our baseline models. Here encoder and decoder parameters are set at the same value.

3.3 Experiments

We explored several ideas and aspects of training during our experiments, which we summarize below.

3.3.1 System Architecture

We tried several configurations of encoder/decoder layers, inspired by previous work such as Araabi and Monz (2020) and van Biljon et al. (2020) which finds that shallower transformers work better in a low-resource scenario. This was the case also for most of our experiments, where smaller models always outperformed the baseline. This holds true even when training on multilingual data. Apart from the baseline, we tested bigger and deeper models, inspired by work such as (Narang et al., 2021; Wei et al., 2022; Wang et al., 2022), on *mni-eng*, which we considered as the "easiest" direction for the models. Preliminary results show degrading performance with the increase of number of encoder layers. We trained on data tokenized with Unigram and a jointly learned vocabulary of 2000, since this was the best performing setup on the validation split. The results of this experiment are given in Table 6, in terms of BLEU score.

One outlier is the translation to and from Assamese, where baseline models, albeit with a dropout of 0.1, outperformed the smaller ones. In these directions, our final systems turned out to be 18/6 models with an embedding dimension of 384 and a feedforward dimension of 1536. However, it should be noted that these final systems were trained in a parallel line of experiments and with a different framework, TorchScale (Ma et al., 2022), which provides further optimization options, such as DeepNorm. Whether the difference in model behavior is due to the difference in training framework is still not clear and could be explored in future work.

3.3.2 Multilingual Training

We trained parent systems on two different multilingual configurations: using all languages in the task, and using only the ones which shared the script. We called these collated dataset *eng-all*, *eng-Beng*, and *eng-Latn* respectively. The intuition here is

	BLEU	ChrF	RIBES	TER	COMET	Place
eng-asm	7.96	27.31	0.31	91.38	0.59	**10/13**
eng-kha	13.90	37.31	0.61	73.99	0.65	**7/11***
eng-lus	20.48	45.60	0.73	61.22	0.68	**9/10**
eng-mni	19.65	53.26	0.66	69.70	0.72	**12/14**
asm-eng	11.29	30.13	0.64	73.39	0.64	**9/13***
kha-eng	12.71	34.55	0.65	78.15	0.56	**6/11***
lus-eng	23.16	43.02	0.72	62.31	0.63	**6/10***
mni-eng	32.18	58.71	0.76	56.35	0.74	**8/14***

Table 5: Summary of the scores of our best submissions reported in the final evaluation. A star (*) denotes the subtasks in which we scored above the organizers' baseline.

enc/dec	BLEU	Increment
4/4	25.89	**+15.41**
6/6	10.48	baseline
8/4	9.52	*-0.96*
12/4	2.95	*-7.53*
16/4	3.08	*-7.4*

Table 6: An example of the effect of changing the depth of the Transformer on the quality of *mni-eng* translation. 4/4 and 6/6 share the same parameters as the final and baseline systems respectively, while other models have an embedding dimension of 384 and a feedforward dimension of 1536.

to leverage script and language relatedness, which we assumed to be present if not for typology, than for script or geographical closeness, in order to obtain better representations of shared subwords and tokens.

We then fine-tuned child systems for each direction, using *eng-Beng* for Assamese and Manipuri, and *eng-Latn* for Khasi and Mizo. *eng-all* was a parent for systems in all directions. We did not specify any language tag or direction for the parent training, since we did not intend to use them for multilingual translation directly. And since the child systems operate only in one direction, we did not need to specify any language tag for fine-tuning either.

Pretraining on all languages proved to be better than standard supervised training for translating into English from Khasi and Mizo, while translation from Manipuri had better performance with the same script parent.

3.3.3 Backtranslation

We experimented with backtranslation in the *eng-mni* direction, by normalizing and deduplicating the provided monolingual data down to around 300k sentence pairs. We then backtranslated the other side with our best available system for the *mni-eng* direction, which had a BLEU score of 32.18. Despite this decent performance, the systems we trained on the backtranslated data, both transformers with 4/4 encoder/decoder layers, embedding dimension of 256 and 384, and feedforward dimension of 1024 and 1536, did not outperform the previous best system. The bigger of the two models had a roughly 2.5 BLEU points on the smaller one, indicating that bigger architectures could have had even better performance. However, we did not test this at this point.

We also tried other back translation approaches, such as Data Diversification (Nguyen et al., 2020), which proved to be effective in the WMT22 Low-resource shared task for Lower/Upper Sorbian and German (Signoroni and Rychlý, 2022b). However, our results using the baseline systems were inconclusive and we decided to explore other approaches.

3.3.4 Word Embeddings Initialization

Previous work (Qi et al., 2018; Edman et al., 2021) showed that using word embeddings to initialize the model's weights improves, sometimes greatly, the performance for low-resource machine translation systems. We tested this in the *eng-mni* direction, training source side word embeddings on the train split with FastText (Bojanowski et al., 2017) in the skipgram setting. While training new baseline systems with the word embedding initialization we observed tiny gains of <0.5 BLEU, however the models converged faster, with 15 to 35 fewer epochs elapsed.

3.3.5 Tokenization Settings

We wanted to explore different settings for the vocabulary size and frequency threshold of the tok-

enizer, as well as for the segmentation algorithm itself, with the objective to find the best settings for each language pair and direction.

Jointly training the vocabulary never resulted in the best system when translating from English, however it gave the best performance for bilingual training of *lus-eng* and *kha-eng*. Nevertheless, these were not the best models overall. With respect to vocabulary size and frequency threshold, in all cases apart from *eng-mni* where we found size 500 and threshold of 200 as best settings, the baselines of 2000 and 100 for these parameters resulted in the best systems. Overall, the picture regarding tokenization and preprocessing settings is not clear and warrants for more investigation.

4 Final Systems

After the experimental phase, we submitted our best performing systems. Table 7 gives their settings and parameters, while Table 5 summarizes the final scores and our placements. Firstly, it should be highlighted that the systems were ranked according to BLEU (Papineni et al., 2002) score. Other metrics, such as ChrF (Popović, 2015), RIBES (Isozaki et al., 2010), TER (Snover et al., 2006), and COMET (Rei et al., 2020), were computed. Looking at the final scores, one can spot several instances in which the metrics do not agree with each other. As a matter of example, the best system for English-Manipuri has 51.96 BLEU, against our twelfth place with 19.65; however our "low-tier" system beats the first one in ChrF (53.26 > 52.61), RIBES (0.66 > 0.51), and COMET (0.72 > 0.57). Recent work has argued for the abandonment of BLEU as a metric of machine translation performance (Kocmi et al., 2021; Mathur et al., 2020; Tan et al., 2015; Sai et al., 2023) in favor of neural metrics which correlate better with human judgements, however this is not always possible when under-resourced languages are involved. While we did not conduct a full and systematic analysis and comparison of the final scores, cases such as the one cited above call for a deeper investigation on automatic evaluation in machine translation.

As our final systems, we obtained roughly two kinds of models: the ones trained on bilingual parallel data, and the ones fine-tuned from a multilingual pair. The former were our best for translating English to the all the Indic languages, and also to translate from Assamese into English. Multilingual pretraining and fine-tuning performed better for the remaining directions, Khasi, Mizo, and Manipuri into English. *kha-eng* and *lus-eng* were fine-tuned from a parent trained on all the parallel dataset, while *mni-eng* was fine-tuned from an Assamese and Manipuri parent. Parent models were trained according to the settings in Table 7, with a patience of 20. fine-tuning was done on only the data for the final translation direction, again with a patience of 20.

Regarding the preprocessing configuration, the settings varied across all the directions. In some cases, such as *eng-kha* and *eng-lus*, sticking to separate source and target vocabulary of size 2000 with a frequency filter of 100 resulted still in the best system. However, for *eng-mni* we found our best system with separate vocabularies of size 500 and a threshold of 200. For multilingual systems, we set the vocabulary size for the Indic side to 750 to try and force the learning of more shared subwords. For the English side, we left the value at 2000. There is no clear winner with respect to segmentation algorithm.

System architecture is the same for all systems, apart from English to and from Assamese. The best architecture was almost always a Transformer with 4 encoder/decoder layers, embedding dimension of 256, feedforward dimension of 1024, and 4 attention heads. For the models involving Assamese, we found that a deeper model of 18 encoder and 6 decoder layers, embedding dimension of 384, feedforward dimension of 1536, and 4 attention heads performed the best.

Other hyperparameters were not investigated extensively, so all of our models were trained with the *adam* optimizer, a learning rate of 1e-3, 4000 warmup updates, a dropout of 0.3, a label smoothing of 0.1, and max tokens for each batch at 16384.

5 Conclusions

This paper describes our experiments and the resulting supervised neural machine translation systems we submitted to WMT23 Low-resource Indic Machine Translation shared task. We trained systems for all directions in the task and experimented with hyperparameter tuning and multilingual training. We did not use transfer learning from pretrained systems, and thus our models were not competitive for some directions. Nonetheless, we argue that our investigation and preliminary analysis on the behavior of different architecture and preprocessing configuration can be useful to other researchers

	eng-asm	eng-kha	eng-lus	eng-mni	asm-eng	kha-eng	lus-eng	mni-eng
training data	bilingual				fine-tuned multilingual			
tokenization	hft			unigram	hft			bpe
src/tgt vocab. size	2000			500	2000	750/2000		
freq. threshold	100			200	100			
enc/dec layers	18/6	4/4			18/6	4/4		
embedding dim.	384	256			384	256		
feedforward dim.	1536	1024			1536	1024		
attention heads	4							
optimizer	adam							
learning rate	1e-3							
warmup updates	4000							
dropout	0.3							
label smoothing	0.1							
max tokens	16384							

Table 7: Summary of the systems for our final submission. The columns give the values for various settings and parameters for preprocessing and training. *bilingual* training denotes a standard supervised training on parallel data, *fine-tuned multilingual* stand for a system fine-tuned on bilingual parallel data, from a parent system trained on more parallel corpora combined. *kha-eng* and *lus-eng* were fine-tuned from a parent trained on all languages, while *mni-eng* parent was trained only on *asm* and *mni* data, which shared the same writing system.

in the field and exposed some interesting points to be explored in future work. Some of our preliminary experiments, such as the use of word embeddings for initialization and backtranslation, did not give the expected results, thus prompting further inquiry.

Limitations and Future Work

As already mentioned above, some instances of disagreement between metrics in the final ranking signal the need for a deeper analysis of the automated evaluation of machine translation. Here, we did not conduct a methodical study on the matter in this instance, this should be the subject for future studies.

The disagreement between metrics notwithstanding, it could be said that overall the performance of our systems was limited. Supervised training showed all its limitations with the small amount of parallel data made available for training. A careful choice of hyperparameters and techniques may ameliorate the situation, but these factors are dependent on the specific dataset involved. Further research must be carried out to uncover clearer connections between the features of the dataset and the choice of parameters and methods to be used. This would cut experimental costs in terms of resources and time, and could lead to better and more efficient models.

However, even if the final systems did not reach competitive levels of performance in some of the cases, our experiments brought up some points that warrant for a deeper investigation. First, the performance of a certain configuration of settings may depend on the framework used for training. The experiments with transformer depth for *mni-eng* contradicts our best systems for Assamese. Whether this discrepancy depends on the languages or on the fact that we used different framework for different translation directions has to be clarified.

Moreover, the connection between dataset and model size has to be investigated further. Assamese worked better with bigger models, even if its dataset was smaller than the multilingual datasets. This goes against the common understanding that a model with fewer parameters is best to deal with fewer data, which will not be enough to train a bigger model. Why this happens only for the Assamese dataset, and not for others, should be better understood.

Ethics Statement

As with any other system trained on real-world data, our models may be biased. These must be taken into account, especially in light of the complex ethnic and religious situation of the region. [2]

Following Lacoste et al. (2019), we report that the experiments and the research that led to the results presented in this paper were conducted

[2]https://www.bbc.com/news/world-asia-india-66086142 (retrieved Aug 31 2023)

on a private server infrastructure consisting of an NVIDIA Tesla T4, A40, and A100 for around 300 hours of training at an efficiency of 0.59 kg/kWh[3] for a total of 44.25 kg CO_2 eq.

References

Ali Araabi and Christof Monz. 2020. Optimizing transformer for low-resource neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3429–3435, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Lukas Edman, Ahmet Üstün, Antonio Toral, and Gertjan van Noord. 2021. Unsupervised translation of German–Lower Sorbian: Exploring training and novel transfer methods on a low-resource language. In *Proceedings of the Sixth Conference on Machine Translation*, pages 982–988, Online. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952, Cambridge, MA. Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

[3]The Czech Republic's country average as reported in https://www.carbonfootprint.com/docs/2018_8_electricity_factors_august_2018_-_online_sources.pdf

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Shuming Ma, Hongyu Wang, Shaohan Huang, Wenhui Wang, Zewen Chi, Li Dong, Alon Benhaim, Barun Patra, Vishrav Chaudhary, Xia Song, and Furu Wei. 2022. TorchScale: Transformers at scale. *CoRR*, abs/2211.13184.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Sharan Narang, Hyung Won Chung, Yi Tay, Liam Fedus, Thibault Fevry, Michael Matena, Karishma Malkan, Noah Fiedel, Noam Shazeer, Zhenzhong Lan, Yanqi Zhou, Wei Li, Nan Ding, Jake Marcus, Adam Roberts, and Colin Raffel. 2021. Do transformer modifications transfer across implementations and applications? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5758–5773, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xuan-Phi Nguyen, Shafiq Joty, Wu Kui, and Ai Ti Aw. 2020. Data diversification: A simple strategy for neural machine translation.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*,

pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ananya B. Sai, Vignesh Nagarajan, Tanay Dixit, Raj Dabre, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2023. Indicmt eval: A dataset to meta-evaluate machine translation metrics for indian languages.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Edoardo Signoroni and Pavel Rychlý. 2022a. HFT: High frequency tokens for low-resource NMT. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 56–63, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Edoardo Signoroni and Pavel Rychlý. 2022b. MUNI-NLP systems for Lower Sorbian-German and Lower Sorbian-Upper Sorbian machine translation @ WMT22. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1111–1116, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Liling Tan, Jon Dehdari, and Josef van Genabith. 2015. An awkward disparity between BLEU / RIBES scores and human judgements in machine translation. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pages 74–81, Kyoto, Japan. Workshop on Asian Translation.

Elan van Biljon, Arnu Pretorius, and Julia Kreutzer. 2020. On optimal transformer depth for low-resource language translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. 2022. Deepnet: Scaling transformers to 1,000 layers.

Daimeng Wei, Zhiqiang Rao, Zhanglin Wu, Shaojun Li, Yuanchang Luo, Yuhao Xie, Xiaoyu Chen, Hengchao Shang, Zongyao Li, Zhengzhe Yu, Jinlong Yang, Miaomiao Ma, Lizhi Lei, Hao Yang, and Ying Qin. 2022. HW-TSC's submissions to the WMT 2022 general machine translation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 403–410, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

NITS-CNLP Low-Resource Neural Machine Translation Systems of English-Manipuri Language Pair

Kshetrimayum Boynao Singh[1], **Ningthoujam Avichandra Singh**[1], **Loitongbam Sanayai Meetei**[1], **Sivaji Bandyopadhyay**[2], and **Thoudam Doren Singh**[1]

[1]Centre for Natural Language Processing (CNLP) & Dept. of CSE, NIT Silchar, India
[2]Dept. of CSE, Jadavpur University, India
{boynfrancis,avichandra0420,loisanayai,sivaji.cse.ju,thoudam.doren}@gmail.com

Abstract

This paper describes the transformer-based Neural Machine translation (NMT) system for the Low-Resource Indic Language Translation task for the English-Manipuri language pair submitted by the Centre for Natural Language Processing in National Institute of Technology Silchar, India (NITS-CNLP) in the WMT 2023 shared task. The model attained an overall BLEU score of 22.75 and 26.92 for the English to Manipuri and Manipuri to English translations respectively. Experimental results for English to Manipuri and Manipuri to English models for character level n-gram F-score (chrF) of 48.35 and 48.64, RIBES of 0.61 and 0.65, TER of 70.02 and 67.62, as well as COMET of 0.70 and 0.66 respectively are reported.

1 Introduction

Our team from Centre for Natural Language Processing at National Institute of Technology Silchar, India (NITS-CNLP) participated in Low-Resource Indic Language Translation task for English-Manipuri language pair in the WMT 2023 shared task (Pal et al., 2023). The shared task involves developing Machine Translation (MT) systems with relatively small parallel datasets. Neural Machine translation (NMT) has been a trending topic for the last few years for translating human languages. Manipuri's MT task is still in its infancy because of the limited resources. Singh and Bandyopadhyay (2011a) conducted a study on supervised statistical methods in which the authors present a convincing study of the impact of morphosyntactic information and dependencies in the context of statistical machine translation. In another work, Singh and Bandyopadhyay (2011b) showed that the expression grounded Statistical Machine translation (SMT) system improves by incorporating verbal features including named entities and reduplicated multiword expressions. Despite the advancement in MT tasks, its investigation in low-resource languages is limited. MT researchers have introduced several approaches to overcome this bottleneck such as data augmentation using back-translation (Sennrich et al., 2016a), multilingual approach (Singh and Singh, 2022a), semi-supervised approach (Cheng et al., 2016; Singh and Singh, 2022b) and exploiting cues from multiple modalities (Gain et al., 2021; Meetei et al., 2023). There are also reports of a comparative study of MT systems on the low resource machine translation focusing on Indian languages such as Assamese (Baruah et al., 2021) and Mizo (Devi et al., 2022; Thangkhanhau and Hussain, 2023).

Driven by the benefits of NMT over traditional MT systems and the encouraging outcomes achieved by NMT in recent times, a study to assess its efficacy in the domain of Indian languages is conducted. Specifically, we have developed and assessed NMT models for translating English to Manipuri and Manipuri to English. The predicted translations are evaluated using automatic evaluation metric and qualitative analysis.

1.1 About the language

Manipuri is the lingua franca of Manipur and has been in existence since 2000 years back till present times with records preserved in the classical cultural heritage of literature. Manipuri is a language of Tibeto-Burman sub-family of the Sino-Tobentan languages family which is locally called as Meeiteilon/Meiteilon (hereon Meiteilon). It is one of the 22 official languages of the India included in the 8th schedule of the Indian constitution[1]. Meiteilon had its original script named Meitei/Meetei Mayek (hereon Meitei Mayek) which was in use up to the 18th century and was replaced later with the Bengali script. However, the wave of revivalist movement

[1]https://rajbhasha.gov.in/en/languages-included-eighth-schedule-indian-constitution

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 967–971
December 6–7, 2023.

emerged later leading to the formation of Meitei Mayek Advisory committee in the year 1973. In 1982, the Government of Manipur announced its decision to include Manipuri in the school education and efforts to revive the Meitei Mayek are still on.

2 System Overview

2.1 Dataset

Language	Sentence	Word	Avg
Eng-Training	21687	390730	18
Man-Training	21687	330319	15
Eng-Validation	1000	16905	16
Man-Validation	1000	14469	14
Eng-Testing	1000	14886	14
Man-Testing	1000	12775	12

Table 1: Statistics of the experimental dataset. (Avg = Average Sentence length

The Manipuri text is written in Bengali script. Statistics of the training dataset are shown in Table 1.

2.2 Data preparation

Training the dataset is pre-processing with subword tokenization. For subword based tokenization we use a source and target BPE of 10000 subword tokens or vocabularies using sentences pieces over the parallel training dataset and applied to the remaining testing and validation dataset. The subword tokenization (Sennrich et al., 2016b) is carried out using the subword-nmt tool [2].

2.3 MT model

Our MT models are trained using OpenNMT toolkit (Klein et al., 2017) and is based on the transformer model (Vaswani et al., 2017).

2.4 Model parameter

Our models are trained for 300000 steps and validated after every 5000 steps. We set the parameter of batch type to tokens and batch size to 2048. The models are trained using Adam optimizer (Kingma and Ba, 2014) with a learning rate of 2 and the dropout set to 0.1. Early stopping mechanism is employed where the training is stopped when the accuracy does not improve for 30 consecutive validations.

In our transformer-based model, each source encoder has 4 layers and decoder also has 4 layers, with a word vector size of 512 and a shared encoder and decoder embedding. We measure the performance of our models by using BLUE (Papineni et al., 2002) and chrF (Popović, 2015).

3 Results and Discussion

In this section, we discuss the experimental results and the performance of models. The reported BLEU score is calculated upon the de-tokenized text using sacrebleu (Post, 2018) while remaining score such as chrF, RIBES, TER and COMET are calculated using tool provided by the organizer[3]. The scores of the system are given in Table 2. The Manipuri to English translation model obtained a BLUE score of 26.92 while English to Manipuri obtained 22.75.

3.1 Qualitative analysis

Automated evaluation metrics such as BLEU measures the precise lexical matches between the translated output and the reference sentences. However, it is inevitable for a natural language to exhibit linguistic variations in terms of synonyms and transformations between active and passive mode of communication. As a result, despite preserving the intended meaning of the source sentence in the translation output, the automated score based on the n-gram match suffers. Manipuri is a language of considerable linguistic diversity and the automated scores for English to Manipuri translations are typically lower than those for Manipuri to English despite the provision of a translation output of acceptable quality. Therefore, we enlist the services of a bilingual native speaker of Manipuri with fluency in English to evaluate the translation outputs for the English to Manipuri task. Table 3 and Table 4 present four randomly selected source sentences from the test set for each of the Manipuri to English and English to Manipuri translation models along with their corresponding output sentences with reference sentences to carry out subjective evaluation.

In Table 3, the difference between the reference and MT results are reported. For Source1, Source2 and Source4, the Manipuri to English MT system outputs (OutputE1, OutputE2 and OutputE4) are close to the reference sentence. In OutputE2, the sentence formation is incorrect where

[2]https://github.com/rsennrich/subword-nmt

[3]http://www2.statmt.org/wmt23/indic-mt-task.html

MT system	BLEU	chrF	RIBES	TER	COMET
en-mni	22.75	48.35	0.61	70.02	0.70
mni-en	26.92	48.64	0.65	67.62	0.66

Table 2: BLEU score and the character n-gram F-score (chrF), RIBES, TER and COMET values of the English→Manipuri (en-mni) and Manipuri→English (mni-en) translation model.

Result	Samples
Source1:	দোক্টর অমা খুদক্তা কৌবীয়ু , নত্ত্রগা অনাবদু হোস্পিটালদা পুনবা এম্বুলেন্স অমা থৌরাং তৌবীয়ু ।
References1:	*call a doctor immediately , or arrange for an ambulance to take the casualty to hospital*
OutputE1:	**send for a doctor immediately , arrange for an ambulance to take the causality to hospital.**
Source2:	ই-বোক্স অসি শিজিন্নবদা য়াম্না লাই অমসুং কমপ্লেন্ত অদুগী অরোন-অথুপ অদু ঙাক্কুনা থম্বদা মতেং পাংগনি ।
References2:	*e-box is very simple to operate and will help to maintain the confidentiality of the complaint.*
OutputE2:	**the e-box will help you protect and will help with regard to confidentiality of the complaint.**
Source3:	মসি বেঙ্গলোর মেথদগী ওন্ন-তৈনবনি ।
References3:	*it is the reverse in bangalore method.*
OutputE3:	**it is to be done from bangalore method.**
Source4:	ভারত্তা অহল ওইরবা মীওইশীংগী মীশীং অসি লেপ্তনা হেনগৎলক্লি ।
References4:	*there has been a steady rise in the population of older persons in india.*
OutputE4:	**the number of older persons has been increasing**

Table 3: Sample input and output of the Manipuri to English MT system.

Result	Samples
Source1:	*encouraging the appropriate government to assume the fullest responsibility for the administration of occupational safety , health and environment at workplace*
References1:	সেফটি , হেলথ এন্ড এনভাইরনমেন্ট এট ৱার্কপ্লেসকী মতাংদা মতিক চাবা লেজিস্লেসন অমা শেমবা
OutputM1:	সেফটি , হেলথ এন্ড এনভাইরনমেন্ট এট ৱার্কপ্লেসকী মতাংদা মতিক চাবা লেজিস্লেসন অমা শেমগৎপা ।
Source2:	*the chairperson of the national authority shall preside over the meetings of the national authority .*
References2:	নেসনেল ওথোরিটিগী চিয়ারপর্সননা নেসনেল ওথোরিটিগী মিটিংশীংগী থৌরম মপু ওইগনি ।
OutputM2:	নেসনেল ওথোরিটিগী চিয়ারপর্সননা নেসনেল ওথোরিটিগী মীটিং অদুগী মীফম পাংথোক্কনি ।
Source3:	*this causes pain .*
References3:	মসিনা নাবা ফাউহল্লি ।
OutputM3:	মসিনা নাবা থোকহল্লিবা মরমশীং
Source4:	*compensation for accredited social health activist*
References4:	এক্রেদিতেদ সোসিএল হেলথ এক্টিভিষ্ট গী কম্পেন্সেসনট
OutputM4:	এক্রেদিতেদ সোসিএল হেলথ এক্টিভিষ্ট গীদমক কম্পেন্সেসন পীবা

Table 4: Sample input and output of the English to Manipuri MT system.

the words such as “maintain” and “operate” are incorrectly translated. In OutputE1, we observe a case where the word “call” is translated to its antonym “send” and “casualty” to “causality”. In OutputE4, “steady rise” is translated as “increasing” which could be considered as a synonym of the phase. Apart from the missing words “in india”, the output sentence preserve the meaning of the source sentence. In OutputE3, the MT output is not able to retain the intended meaning of the source sentence.

The Table 4 shows the results of the MT system for translating English to Manipuri. The OutputM1, OutputM3 and OutputM4 give a close meaning to the reference sentence. In OutputM1, the word “শেমবা” (meaning “build”) is translated to its infinitive form of the verb “শেমগৎপা” (meaning “to build”). The word “causes” has multiple translations in Manipuri such as “ফাউহল্লি”, “থোকহল্লিবা”and “মরমশীং” which are used in different context. In OutputM3, we observe that the translations of the word “causes” is repeated showing the challenges of the MT model in translating such words. In OutputM4, we observe a case where a word as multiple translation in Manipur but can be used in the same context. The word “for” can be translated as “গী” or “গীদমক” in Manipuri. Apart from the extra verb “পীবা” (meaning “give”), OutputM4 is grammatically correct despite not having a perfect n-gram match. In the case of OutputM2, the sentence is observe to have a poor adequacy with the incorrect translation for the word “chairperson” but the structure of sentences is well formed and grammatically correct.

4 Conclusion

Enabling MT for low-resource languages poses several challenges due to the lack of parallel resources available for training. In this work, we report the performance of the MT systems trained on low resource setting for English to Manipuri and Manipuri to English using a transformer-based encoder and decoder architecture. The automatic evaluation shows that English to Manipuri MT system achieved 22.75 BLEU and Manipuri to English MT system achieved 26.92 BLEU. The automated scoring mechanism is inadequate in capturing the linguistic nuances of the morphologically complex Manipuri language which requires the use of multiple references. Based on the subjective evaluation, we observed that the translation quality is deemed satisfactory and fluent in some cases, given the relatively small size of the dataset and the utilization of a single test reference.

Limitations

Translation model performs better for short sentences as compared to the longer sentences. There are several out of vocabulary words due to the fact that the model is built on a constraint environment.

Acknowledgements

This work is sponsored by MEITY Ref. No. 11(1)/2022-HCC(TDIL)-Part(4). We also acknowledge CNLP (Centre for Natural Language Processing) and Department of Computer Science and Engineering at National Institute of Technology Silchar for providing and giving access to the computing facilities.

References

Rupjyoti Baruah, Rajesh Kumar Mundotiya, and Anil Kumar Singh. 2021. Low resource neural machine translation: Assamese to/from other indo-aryan (indic) languages. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 21(1):1–32.

Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Semi-supervised learning for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1965–1974.

Chanambam Sveta Devi, Bipul Syam Purkayastha, and Loitongbam Sanayai Meetei. 2022. An empirical study on english-mizo statistical machine translation with bible corpus. *International journal of electrical and computer engineering systems*, 13(9):759–765.

Baban Gain, Dibyanayan Bandyopadhyay, and Asif Ekbal. 2021. Experiences of adapting multimodal machine translation techniques for hindi. In *Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021)*, pages 40–44.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Loitongbam Sanayai Meetei, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2023. Exploiting multiple correlated modalities can enhance low-resource machine translation quality. *Multimedia Tools and Applications*, pages 1–21.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Sandeep Kumar Dash, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, and Pankaj Kundan Dadure. 2023. Findings of the wmt 2023 shared task on low-resource indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Salam Michael Singh and Thoudam Doren Singh. 2022a. An empirical study of low-resource neural machine translation of manipuri in multilingual settings. *Neural Computing and Applications*, 34(17):14823–14844.

Salam Michael Singh and Thoudam Doren Singh. 2022b. Low resource machine translation of english–manipuri: A semi-supervised approach. *Expert Systems with Applications*, 209:118187.

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2011a. Bidirectional statistical machine translation of manipuri english language pair using morpho-syntactic and dependency relations. *International Journal of Translation*, 23(1):115–137.

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2011b. Integration of reduplicated multiword expressions and named entities in a phrase based statistical machine translation system. In *Proceedings of 5th international joint conference on natural language processing*, pages 1304–1312.

Haulai Thangkhanhau and Jamal Hussain. 2023. Construction of mizo–english parallel corpus for machine translation. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

IACS-LRILT: Machine Translation for Low-Resource Indic Languages

Dhairya Suman[1⋆] and **Atanu Mandal**[2†] and **Santanu Pal**[3‡] and **Sudip Kumar Naskar**[4†]
⋆Indian Association for the Cultivation of Science, Kolkata, India
†Jadavpur University, Kolkata, India
‡Wipro AI Lab, London, UK
{[1]dhairyasuman, [2]atanumandal0491, [3]santanu.pal.ju, [4]sudip.naskar}@gmail.com

Abstract

Even though, machine translation has seen huge improvements in the the last decade, translation quality for Indic languages is still underwhelming, which is attributed to the small amount of parallel data available. In this paper, we present our approach to mitigate the issue of the low amount of parallel training data availability for Indic languages, especially for the language pair English-Manipuri and Assamese-English. Our primary submission for the Manipuri-to-English translation task provided the best scoring system for this language direction. We describe about the systems we built in detail and our findings in the process.

1 Introduction

The ability to overcome linguistic barriers has emerged as the most critical issue in a society that is becoming increasingly interconnected. These linguistic barriers can be eliminated enabling effective communication among various linguistic communities, machine translation (MT) systems are not only capable of translating common languages but also less widely spoken or even endangered languages, ensuring that even marginalized communities can participate in the global conversation. The use of machine translation for regional Indian languages is both an intriguing and challenging application. India has an intricate mix of languages and dialects spoken all over its broad territory, making it a linguistically diverse nation (Mandal et al., 2021). Despite being culturally stimulating, this diversity poses substantial obstacles to effective communication. By automating the translation process and opening up content to speakers of different regional Indian languages, machine translation presents a viable remedy.

Due to deep learning, neural networks, and natural language processing developments, machine translation technology has made significant strides in recent years (Slocum, 1985). However, there are particular difficulties that must be overcome in order to adapt these technologies to the intricate linguistic features of Indian languages (Pal et al., 2013a). These difficulties include, among other things e.g., multi-word expressions (Pal et al., 2013b), the complexity of morphology, syntactic changes, and the scarcity of parallel training data (Pal, 2018). The challenge of producing accurate and relevant translations is further complicated by the requirement to preserve cultural nuances and context-specific meanings (Appicharla et al., 2023).

However, the translation problem for Indian regional languages is compounded by:

- **Morphological complexity:** Indian languages often exhibit rich morphology, leading to variations in word forms and sentence structures.
- **Low-resource languages:** Limited parallel training data is available for many Indian language pairs, leading to challenges in training accurate translation models.
- **Cultural and context preservation:** Accurate translation must account for context-specific meanings, idiomatic expressions, and cultural nuances.

So working on Indic languages has the challenge of designing translation models and techniques that address these complexities and constraints while achieving high-quality translations between Indian regional languages, contributing to effective cross-lingual communication and content accessibility in India's diverse linguistic landscape.

2 Related Work

Parul and Garg (2022) provides a survey of different approaches to Machine Translation (MT) for Indian languages, including Rule-based Machine Translation (RBMT), Corpus-based Machine

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 972–977

Translation (CBMT), and Neural Machine Translation (NMT). Researcher (Parul and Garg, 2022) highlights the initial slow progress in MT research and the subsequent popularity of NMT. The paper emphasizes that while there has been significant research on MT for top-level languages, there is a scarcity of research for low-level languages spoken by fewer people. It discusses the use of different MT models, such as Anusaarka for direct MT, AnglaHindi for Interlingual translation, and CBMT for translation using stored data corpus.

Jha et al. (2023) presents the development and evaluation of a multilingual neural machine translation system for Indian languages using the mT5 transformer. The system was trained on the modified Asian Language Treebank multilingual dataset to translate text between English, Hindi, and Bengali. The system achieved acceptable Bilingual Evaluation Understudy (BLEU) scores, with the English-to-Bengali system achieving a maximum BLEU score of 49.87 and the Bengali-to-English system achieving an average BLEU score of 42.43. Jha et al. (2023) claims that the field of Natural Language Processing (NLP) research in low-resource languages has been expanding rapidly, with transformers being the latest state-of-the-art systems.

Jayanthi et al. (2020) states that India is a multicultural and multilingual country, with a large number of regional languages. English is provided as the second extra official language in India, but its usage is limited, leading to a communication gap. Machine translation can help minimize this gap by translating languages. Jayanthi et al. (2020) focuses on translating Indic languages, specifically Telugu, using a sequence-to-sequence framework with an encoder-decoder attention mechanism of neural machine translation. The proposed framework aims to convert the Telugu language into English and vice versa. Their approach framework was trained using a Telugu parallel corpus and achieved good accuracy. It overcomes the limitation of reduced accuracy when faced with unknown words by using an attention mechanism. As per the author, the sequence-to-sequence model used in this paper allows for the conversion of the native language into the desired language, and the attention mechanism helps handle rare words.

S. and Bhattacharyya (2020) claims the use of Indowordnet helped handle ambiguity during translation and improved the performance of the machine translation systems. The author presents a comparative study of 440 phrase-based statistically trained models for 110 language pairs across 11 Indian languages and also discusses the principles followed in constructing the synsets, such as the minimality principle, coverage principle, and replaceability principle.

Research involving Indian languages is not very common due to the scarcity of parallel corpora. Baruah et al. (2014) using Statistical Machine Translation (SMT) with a small corpus (2,500 sentences), the Assamese-English bidirectional MT system for Assamese to English and English to Assamese obtained BLEU scores of 9.72 and 5.02, respectively. Das and Baruah (2014) investigated and reported a BLEU score of 11.32 for Assamese to English using SMT using 8,000 Tourism domain parallel sentences.

3 Method

3.1 Problem Definition

Given a source sentence in an Indian regional language, represented as $S = \{s_1, s_2, \ldots, s_n\}$, and a target sentence in a different Indian regional language or English, represented as $T = \{t_1, t_2, \ldots, t_m\}$, the objective of machine translation for Indian regional languages is to find the optimal translation function f that maximizes the translation quality while considering linguistic nuances, morphological complexities, and contextual information:

$$f^\star = \text{argmax}\, f(P(T \mid S)) \tag{1}$$

In equation 1, $f^\star$ represents the optimal translation function that produces the highest probability of the target sentence given the source sentence. $P(T \mid S)$ is the conditional probability of the target sentence T given the source sentence S, which is modelled using statistical or neural machine translation approaches. $S = \{s_1, s_2, \ldots, s_n\}$ denotes the sequence of words in the source sentence. $T = \{t_1, t_2, \ldots, t_m\}$ denotes the sequence of words in the target sentence. n is the length of the source sentence, and m is the length of the target sentence.

3.2 Dataset Description

Table 1 represents the Datasets for the language pair of Assamese-English and Manipur-English language pair in the WMT 2023 IndicMT[1] shared

[1] http://www2.statmt.org/wmt23/indic-mt-task.html

Language Pair	Train	Validation	Test
Assamese-English	50,000	2,000	2,000
English-Manipuri	21,686	1,000	1,000

Table 1: Dataset statistics for Workshop on Machine Translation (WMT) 23

task. As per the organizers' guidelines, no additional parallel data was allowed for training with only constrained submissions.

3.3 Experimental Setup

IndicBART (Dabre et al., 2022) and mbart-large-50 (Tang et al., 2020) have been adjusted for the bidirectional Assamese-English and English-Manipuri language pairs in our suggested study. We fixed the source and target lengths in both scenarios to "128". With batch sizes of "16" and "8", respectively, and learning rates of "2×10^{-5}" for both scenarios, we improved our suggested IndicBART and mbart-large-50 models, We applied weight decay of "0.01" for both scenarios.

3.4 Corpus Pre-processing

We used IndicBART (Dabre et al., 2022) developed by AI4Bharat[2] for some of the models. Using IndicBART for Indic languages other than Hindi or Marathi requires the language to be transliterated into the Devanagari script. Hence, we had to transliterate the data given into the Devanagari script to use those models.

3.5 Experiments

3.5.1 Bidirectional Assamese-English Language Pair

We first experimented by using IndicTrans (Ramesh et al., 2022) from AI4Bharat to get the responses on the Validation Set provided, but the BLEU scores on the same were unsatisfactory. We experimented by finetuning IndicBART from AI4Bharat on the Training Set and evaluating the responses on the given Validation Set. This gave us better results so we decided that these responses would be our Primary Submissions. IndicBART is a multilingual, sequence-to-sequence pre-trained model focusing on Indic Languages and English. Currently, it supports 11 Indian languages, Assamese, Bengali, Gujarati, Hindi, Marathi, Odia, Punjabi, Kannada, Malayalam, Tamil, and Telugu based on mBART (Liu et al., 2020) architecture.

[2] https://ai4bharat.iitm.ac.in/

We used the transliteration module from the IndicNLP library (Kunchukuttan, 2020) for transliterations from Assamese to Devanagari, an example is shown in Figure 1.

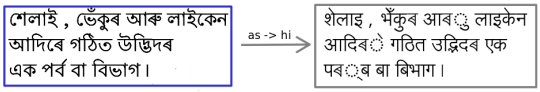


Figure 1: Transliteration from Assamese to Hindi

These experiments are discussed below:

- **Primary Submission**
  We took the training data and fine-tuned it on IndicBART for the translation settings from Assamese to English. This model gave good BLEU scores on the Validation set hence, this model was selected as the Primary System.

- **Contrastive - 1**
  Here, it was considered that since Assamese and Bengali share linguistic similarities, it may be that IndicBART fine-tuned on the training dats but this time for translations from English-Bengali, did give results, surprisingly similar to the Primary System

- **Contrastive - 2**
  Here we used IndicTrans from AI4Bharat, the translator was built and the responses on the Test Set were calculated. Note, that for this system no Transliteration was required.

For the models that used IndicBART, we had to transliterate the data from Assamese to Hindi using the IndicNLP transliterator. Moreover, the responses generated by these models, when the target language was Assamese also had to be back-transliterated from Hindi to Assamese for the evaluation of the Validation Set.

3.5.2 Bidirectional English-Manipuri Language Pair

Since resources available for the Manipuri language are very scarce, we decided to use existing models available for Bengali and Assamese. This was because Manipuri shares its script with Assamese and Bengali, so even with morphological differences the models gave good scores for Manipuri. We used mbart-large-50 (Tang et al., 2020) from Facebook and IndicBART by AI4Bharat.

For the language pair English-Manipuri there were no existing transliteration tools that we found,

Framework	BLEU	ChrF	RIBES	TER	COMET
English-to-Assamese					
Primary	34.82	56.58	0.87	55.10	0.77
Contrastive-1	34.71	56.59	0.87	54.75	0.78
Benchmark	8.57	25.24	0.44	86.14	0.59
Contrastive-2	6.57	39.71	0.45	86.26	0.79
Assamese-to-English					
Primary	66.36	75.88	0.93	37.44	0.84
Contrastive-1	66.33	75.88	0.93	37.38	0.84
Contrastive-2	23.19	48.42	0.61	71.79	0.75
Benchmark	11.28	28.70	0.53	83.10	0.56
English-to-Manipuri					
Primary	25.78	49.94	0.84	60.43	0.71
Contrastive-1	25.82	49.93	0.84	60.57	0.71
Benchmark	21.58	45.97	0.61	69.76	0.69
Contrastive-2	9.69	40.45	0.54	81.18	0.67
Manipuri-to-English					
Primary	69.75	78.16	0.94	32.08	0.84
Contrastive-1	69.75	78.16	0.94	32.10	0.84
Benchmark	24.86	46.37	0.64	70.26	0.63
Contrastive-2	22.10	48.03	0.63	72.19	0.70

Table 2: Results of Primary, Contrastive-1, and Contrastive-2 submissions evaluated on Benchmark results for the language pair Assamese-English and English-Manipuri.

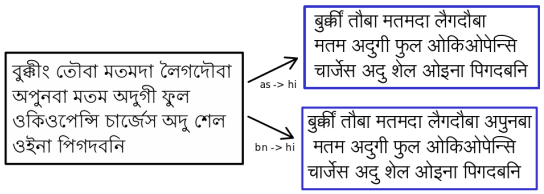


Figure 2: Transliteration from Manipuri to Hindi

but it was thought that, since Manipuri has script similarities with Bengali and Assamese we can experiment with transliteration tools from Bengali and Assamese to Hindi with the expectation for good results and it turns out it does give good results. For this task too we used the transliteration tools from the IndicNLP library, an example is shown in Figure 2.

We discussed in detail about these experiments as follows:

- **Primary Submission**
  The data was first transliterated into Hindi using the transliteration from Bengali to Hindi, then we finetuned IndicBART on the Training Data and evaluated the responses given for the Validation Set. This model gave the highest score on the Validation Set and hence, was picked as the Primary model.

- **Contrastive - 1 Submission**
  This model was similar to the Primary model, but instead of the transliteration and Translation settings, Bengali was the Indic language instead of Assamese.

- **Contrastive - 2 Submission**
  For this model, we fine-tuned mbart-large-50 with the Bengali-English configuration. This model gave a lesser score on the validation set than the models discussed before, even though this was a larger model.

Similar, to the Primary and Contrastive - 1 system for Task 1, responses from the models that used IndicBART had to be back-transliterated from Hindi to the Indic language, when the Indic language was the target language.

3.6 Post-processing

Along with the back-transliteration that was required for the models using IndicBART when the target language was the Indic language. We also

had to do some post-processing of the responses received, we saw that often the responses had random Chinese characters and emoticons in the responses. The emoticons were chalked up to encoding errors while saving the responses to a text file, on the other hand, the Chinese characters were something that we think were errors because of the model itself. These noisy characters were manually removed to ensure that they don't affect the accuracy.

4 Results and Analysis

Table 2 lists the findings of our experiments. We list our observations here:

- As we discussed in section 3.6 we believe that there might be noise in the responses saved that we missed or couldn't manually find, which can contribute to a lesser score even though the translations are accurate.
- We also believe that there might be some issues in translation because of transliteration problems while back-transliterating we often came across responses that still had some words in Hindi. Due to this we also believe that there might have been errors in transliteration from the Assamese/Manipuri to Hindi.
- For task 4, we also consider that the transliteration and translation models used were configured to Assamese and Bengali, so even though the models were fine-tuned on the data but still we assume that because of the morphological differences, there might be gaps in the understanding and generating of language by the model.
- An interesting observation that can be made is that there exists a large gap in the scores for when English is the target language and when the target language is the Indic language. This error can be attributed to the model understanding the target languages morphologically well, but not being able to generate the language that well.

5 Conclusion and Future Work

In this paper, We discussed the models and procedures our team used for the language pairs Assamese-English and English-Manipuri. According to our experiments, we claimed that Using language models like IndicBART and mbart-large-50 results in improvement for the low-resourced individual languages results. We hope that this will enable us to develop more precise and superior translation models for languages and domains with limited resources specially for Indian Languages where there is a presence of large language diversity. We also believe that, as seen with Manipuri, a language with very few resources for processing we can use languages close and similar to it to aid in its processing and create a better way of processing those low-resource languages. In future, we will include our models in online post-editing platforms (Pal et al., 2016; Nayak et al., 2015; Vela et al., 2019).

Acknowledgements

This research was supported by the TPU Research Cloud (TRC) program, a Google Research initiative and funded by the 'VIDYAAPATI: Bidirectional Machine Translation Involving Bengali, Konkani, Maithili, Marathi, and Hindi' under the Project titled 'National Language Translation Mission (NLTM): BHASHINI'.

References

Ramakrishna Appicharla, Baban Gain, Santanu Pal, and Asif Ekbal. 2023. A case study on context encoding in multi-encoder based document-level neural machine translation. *arXiv preprint arXiv:2308.06063*.

Kalyanee Baruah, Pranjal Das, Abdul Hannan, and Shikhar Sarma. 2014. Assamese-english bilingual machine translation. *International Journal on Natural Language Computing*, 3.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh M. Khapra, and Pratyush Kumar. 2022. Indicbart: A pre-trained model for natural language generation of indic languages. In *Findings of the Association for Computational Linguistics*.

Pranjal Das and Kalyanee Kanchan Baruah. 2014. Assamese to english statistical machine translation integrated with a transliteration module. *International Journal of Computer Applications*, 100:20–24.

N Jayanthi, Aluri Lakshmi, Ch Suresh Kumar Raju, and B Swathi. 2020. Dual translation of international and indian regional language using recent machine translation. In *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, pages 682–686.

Abhinav Jha, Hemprasad Yashwant Patil, Sumit Kumar Jindal, and Sardar M N Islam. 2023. Multilingual indian language neural machine translation system

using mt5 transformer. In *2023 2nd International Conference on Paradigm Shifts in Communications Embedded Systems, Machine Learning and Signal Processing (PCEMS)*, pages 1–5.

Anoop Kunchukuttan. 2020. The IndicNLP Library. https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Atanu Mandal, Santanu Pal, Indranil Dutta, Mahidas Bhattacharya, and Sudip Kumar Naskar. 2021. Is attention always needed? a case study on language identification from speech. *ArXiv*, abs/2110.03427.

Tapas Nayak, Sudip Kumar Naskar, Santanu Pal, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. 2015. Catalog: New approaches to tm and post editing interfaces. In *Proceedings of the Workshop Natural Language Processing for Translation Memories*, pages 36–42.

Santanu Pal. 2018. A hybrid machine translation framework for an improved translation workflow.

Santanu Pal, Mahammed Hasanuzzaman, Sudip Kumar Naskar, and Sivaji Bandyopadhyay. 2013a. Impact of linguistically motivated shallow phrases in pb-smt. In *ICON 2013*. https://www.researchgate.net/publication ....

Santanu Pal, Sudip Kumar Naskar, and Sivaji Bandyopadhyay. 2013b. Mwe alignment in phrase based statistical machine translation. In *Proceedings of the XIV Machine Translation Summit*, pages 61–68.

Santanu Pal, Sudip Kumar Naskar, Marcos Zampieri, Tapas Nayak, and Josef van Genabith. 2016. CATaLog online: A web-based CAT tool for distributed translation with data capture for APE and translation process research. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 98–102, Osaka, Japan. The COLING 2016 Organizing Committee.

Parul and Kamal Deep Garg. 2022. Machine translation system for indian language: Survey. In *2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC)*, pages 468–473.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Sreelekha S. and Pushpak Bhattacharyya. 2020. Indowordnet's help in indian language machine translation. *AI Soc.*, 35(3):689–698.

Jonathan Slocum. 1985. A survey of machine translation: Its history, current status and future prospects. *Computational Linguistics*, 11(1):1–17.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Mihaela Vela, Santanu Pal, Marcos Zampieri, Sudip Naskar, and Josef van Genabith. 2019. Improving CAT tools in the translation workflow: New approaches and evaluation. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 8–15, Dublin, Ireland. European Association for Machine Translation.

IOL Research Machine Translation Systems for WMT23 Low-Resource Indic Language Translation Shared Task

Wenbo Zhang, Zeyu Yan, Qiaobo Deng, Jie Cai, and **Hongbao Mao**
Transn IOL Research, Wuhan, China

Abstract

This paper describes the IOL Research team's submission systems for the WMT23 low-resource Indic language translation shared task. We participated in 4 language pairs, including en↔as, en↔mz, en↔kha, en↔mn. We use transformer based neural network architecture to train our machine translation models. Overall, the core of our system is to improve the quality of low resource translation by utilizing monolingual data through pre-training and data augmentation. We first trained two denoising language models similar to T5 and BART using monolingual data, and then used parallel data to fine-tune the pretrained language models to obtain two multilingual machine translation models. The multilingual machine translation models can be used to translate English monolingual data into other multilingual data, forming multilingual parallel data as augmented data. We trained multiple translation models from scratch using augmented data and real parallel data to build the final submission systems by model ensemble. Experimental results show that our method greatly improves the BLEU scores for translation of these four language pairs.

1 Introduction

This paper describes our submissions for the WMT23 low-resource Indic language translation shared task. We participated in 4 language pairs, including English↔Assamese (en↔as), English↔Mizo (en↔mz), English↔Khasi (en↔kha), and English↔Manipuri (en↔mn).

Our core approach is based on denoising language model pre-training(Devlin et al., 2019; Lample and Conneau, 2019; Song et al., 2019; Raffel et al., 2019; Lewis et al., 2020) and back-translation(Sennrich et al., 2016a) based data augmentation. Neural machine translation methods are almost the first choice for implementing translation systems at present, but they have certain requirements on the amount of parallel corpora. Low-resource or even zero-resource neural machine translation has been a daunting challenge due to the lack of adequate parallel corpora. Pre-training methods are popular solutions for low-resource cases. When the model parameter scale is large enough and there is enough training data, this method can even perform well in zero resource situations. For the machine translation task, as early as around 2019, XLM(Lample and Conneau, 2019) and MASS(Song et al., 2019) were able to build unsupervised machine translation systems with near-supervised effects using only monolingual data. Now, more advanced pre-training methods like BART(Lewis et al., 2020) and T5(Raffel et al., 2019) are popular choices for training machine translation models in low-resource situations. Therefore, in this paper, referring to the training methods of BART and T5, we trained a T5-style pre-training model and a BART-style pre-training model from scratch using monolingual data. Back-translation is a commonly used method in the field of machine translation. Whether it is low-resource, medium-resource or high-resource, this approach can almost help the model to obtain further improvements on the original basis. Therefore, we also use back-translation to help us further improve the translation quality.

The layout of the subsequent paper is as follows: In Section 2 We introduce the data source and processing strategy; In Section 3 we describes the implementation process of our translation systems; In Section 4 we describe the experimental settings and results; Finally, the conclusion is drawn in Section 5.

2 Data

2.1 Data Source

Bilingual corpus We just used the official en↔as, en↔mz, en↔kha, and en↔mn parallel data(Pal

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 978–982
December 6–7, 2023.

Data	en↔as	en↔mz	en↔kha	en↔mn
Bilingual Data	49808	49575	23996	20990

Table 1: Statistics of bilingual data

Data	en	as	mz	kha	mn
Monolingual Data	60598321	2206328	1864322	178036	298072

Table 2: Statistics of monolingual data

et al., 2023).

Monolingual corpus of Indic languages We also used only official monolingual data for Assamese, Mizo, Khasi and Manipuri.

English monolingual corpus Since the official did not provide English monolingual data, we obtained English monolingual data from the WMT23 general task. Specifically, we used the English side of bilingual data (English↔German and English↔Japanese) in the WMT23 general task as English monolingual data.

2.2 Data Preprocessing

For English monolingual data, we first filter out noisy sentences according to following rules:

- Remove invisible characters.
- Remove sentences containing too more than 300 words or more than 1000 characters or less than 3 characters.
- Remove English sentences containing words exceeding than 40 characters.
- Remove sentences that contain too many punctuation marks.
- Remove sentences that contain repeated substrings, which refers to a string composed of a single character that repeats more than 10 times, or two or more character that repeat more than 5 times.
- Remove sentences that contain HTML tags.
- Convert full-width characters to half-width characters.
- Remove duplicated sentence pairs.

Since all the officially provided data have been tokenized, we used the Moses scripts[1] to do tokenization for English monolingual data. Then we use an n-gram language model trained with KenLM(Heafield, 2011)[2] to calculate the perplexity of English monolingual data and remove sentences with high perplexity(more than 10 000). We just did deduplication for the official data, because the size of the official data is relatively small and the quality is high enough. The amount of data after processing is shown in Table 1 and 2.

We used the Sentencepiece(Kudo and Richardson, 2018) tool to train a multilingual BPE(Sennrich et al., 2016b) model for subword segmentation. Its training data includes all official training data and 2.5 million random samples from English monolingual data. The vocabulary size is set to 48 000.

3 System Overview

We chose Transformer(Vaswani et al., 2017) with pre-norm as our base translation model. In general, our procedure for improving the quality of low-resource translations is divided into two phases, an improvement phase based on pre-training methods and an improvement phase based on data augmentation. Instead of using the pre-trained model to initialize the parameters of the translation model, the pre-training phase merely provides synthetic data for the data augmentation phase, which means that the translation model in the data augmentation phase is trained from scratch. In addition to this, we also used model ensemble in the final submissions.

3.1 Pre-training

The pre-training phase is divided into two steps. In the first step, pre-training for the denoising autoencoder tasks are performed using monolingual data. In the second step, the pre-trained models are fine-tuned using bilingual data. We trained two denoising pre-training models, namely the T5-style(Raffel et al., 2019) model and the BART-style(Lewis et al., 2020) model. The training details

[1]https://github.com/moses-smt/mosesdecoder/

[2]https://github.com/kpu/kenlm

Original sentence	Since their articles appeared , the price of gold has moved up still further .
T5-style input sentence	Since their articles appeared , gold has moved up still further
T5-style target sentence	the price of .
BART-style input sentence	Since their of gold has up still **moved** further .
BART-style target sentence	Since their articles appeared , the price of gold has **moved** up still further .

Table 3: Examples of T5-style and BART-style training data

of the two models are as follows.

As shown in Table 3, Both T5-style and BART-style models are trained by recovering original sentences from corrupted sentences, which are produced by randomly replacing some fragments in the sentences with the mark. The most important difference is that the T5-style target sentence, that is, the label contains only the replaced part, while the BART-style label is the entire original sentence. Another difference is that in this paper we also randomly swap the two words that are not masked in BART-style input sentences. For both models, the proportion of replaced words is 0.15, and the length of replaced segments is 3. We randomly swap words in BART-style input sentences with a probability of 0.5.

We used monolingual data containing 5 languages to train the pre-training models, and then fine-tuned the pre-trained models using parallel corpora containing 4 language pairs in 8 translation directions. In order to keep the number of all languages balanced, we only used 3 million additional English monolingual data at this phase.

3.2 Data Augmentation

The pre-training phase is also divided into two steps, pre-training on synthetic data and fine-tuning on the real bilingual data. We employed the approach inspired by the back-translation(Sennrich et al., 2016a) and Zan et al. (2022) to generate synthetic data. Since we planed to train a multilingual translation model, in order to share knowledge across multiple languages, the synthetic data we generated contains 5 languages and 20 translation directions. In detail, by beam search, we translated an English monolingual sentence into 4 other languages, where any two sentences in different languages are also aligned as they are both translated from the same English sentence. To ensure the quality of the synthesized data, we also calculated the translation perplexity score from Indic languages to English direction via a multilingual translation model from pre-training phase and removed sentence pairs with high perplexity scores. For data diversity, we used both T5-style and BART-style pre-trained models to generate synthetic data, and leveraged the other model to compute the perplexity score, for example, the data generated by the T5-style pre-trained model is scored using the BART-style pre-trained model.

3.3 Model Ensemble

A well-known model ensemble trick is to increase the diversity between different models. However, we did not train multiple translation models from scratch due to time and computational resource constraints. Instead, we fine-tuned the three models, many-to-many, one-to-many, and many-to-one, based on model trained on synthetic data, and then selected the many-to-many and one-to-many or many-to-one models to complete the final submission by model ensemble.

4 Experiments

4.1 Experiment Settings

All of our translation models were implemented based on fairseq(Ott et al., 2019) and trained on 8 NVIDIA A100 GPUs. During training, we used the Adam(Kingma and Ba, 2014) optimizer with $\beta 1 = 0.9, \beta 2 = 0.98$, the learning rate scheduling strategy of inverse sqrt, the number of warmup step set to 4000, the maximum learning rate set to 0.0005 and FP16 to accelerate the training process.

We trained three models, Many2Many(M2M), One2Many(O2M), Many2One(M2O), with 12-encoder, 12-decoder transformer-big model as baselines. They were trained only on a real parallel corpus, with a batch size set at 13,000 tokens. For the models in the Pre-training phase, we used the same model structure as the baselines but with a batch size of 1 million. For the models in the data augmentation phase, we changed the number of layers of models to 10, and the embedding size to 1536.

System	en→as	en→kha	en→mni	en→mz	as→en	kha→en	mni→en	mz→en
O2M Baseline	5.1	11.8	9.1	15.0	-	-	-	-
M2O Baseline	-	-	-	-	14.3	10.6	19.8	18.8
M2M Baseline	7.0	14.8	13.4	19.2	15.9	11.7	23.3	20.6
BART-style Pre-training	11.4	19.3	20.1	25.2	22.4	15.1	35.4	26.5
T5-style Pre-training	12.0	19.6	21.5	26.3	23.6	16.4	35.6	26.9
O2M Data Augmentation	13.0	21.3	23.3	27.4	-	-	-	-
M2O Data Augmentation	-	-	-	-	28.2	20.1	42.1	31.8
M2M Data Augmentation	12.8	21.0	23.4	27.3	25.2	18.0	40.6	29.1
Model Ensemble	13.4	21.6	23.9	27.8	28.6	20.8	42.9	32.4

Table 4: BLEU scores of all translation direction on validation sets

4.2 Results

All experiments were evaluated using the sacrebleu(Post, 2018) tool to calculate BLEU(Papineni et al., 2002) scores on the official validation sets, and we did not detok before calculating the BLEU scores. We used beam search with beam size=5 to decode all models and the results are shown in Table 4.

According to Table 4, it can be seen that the many-to-many baseline preforms better than one-to-many and many-to-one. I believe this is because the parallel corpus size is too small where the many-to-many model can share knowledge across different languages. Both BART-style and T5-style pre-training significantly improved BLEU scores in all directions, with T5-style slightly better than BART-style. All translation directions are further improved after data augmentation. When English is the source language, the improvement is small, and when English is the target language, the improvement is larger. This is because this phase is mainly based on a large amount of real English monolingual data. The one-to-many and many-to-one models perform equally or better than the many-to-many model at this phrase, as there is no longer a severe lack of linguistic knowledge. Finally, the model ensemble helps the system to obtain further improvements.

5 Conclusion

In this paper, we describe IOL Research's submission to the WMT2023 low-resource Indic language translation shared task. We participated in four sub-tasks with a total of eight translation directions. Our system mainly improves the translation quality of these languages in the low-resource case through pre-training and data augmentation. Experimental results show that we achieved large improvements in all directions.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197.

DiederikP. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv: Learning,arXiv: Learning*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv: Computation and Language,arXiv: Computation and Language*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Sandeep Kumar Dash, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, and Pankaj Kundan Dadure. 2023. Findings of the wmt 2023 shared task on low-resource indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and PeterJ. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv: Learning,arXiv: Learning*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. *arXiv: Computation and Language,arXiv: Computation and Language*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Changtong Zan, Keqin Peng, Liang Ding, Baopu Qiu, Boan Liu, Shwai He, Qingyu Lu, Zheng Zhang, Chuang Liu, Weifeng Liu, Yibing Zhan, and Dacheng Tao. 2022. Vega-MT: The JD explore academy machine translation system for WMT22. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 411–422, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Trained MT Metrics Learn to Cope with Machine-translated References

Jannis Vamvas[2*] **Tobias Domhan**[1] **Sony Trenous**[1]
Rico Sennrich[2] **Eva Hasler**[1]
[1]Amazon AI Translate, Berlin
[2]University of Zurich
{vamvas,sennrich}@cl.uzh.ch, {domhant,trenous,ehasler}@amazon.com

Abstract

Neural metrics trained on human evaluations of MT tend to correlate well with human judgments, but their behavior is not fully understood. In this paper, we perform a controlled experiment and compare a baseline metric that has not been trained on human evaluations (*Prism*) to a trained version of the same metric (*Prism+FT*). Surprisingly, we find that Prism+FT becomes more robust to machine-translated references, which are a notorious problem in MT evaluation. This suggests that the effects of metric training go beyond the intended effect of improving overall correlation with human judgments.

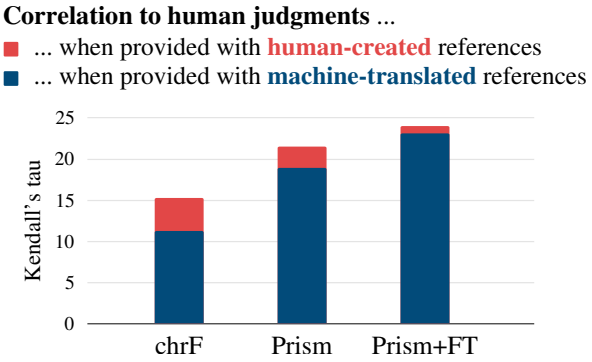


Figure 1: Metrics for MT quality have a lower segment-level correlation with human judgments when provided with machine-translated references. However, trained metrics, such as our Prism+FT, become more robust to the use of machine translations as references.

1 Introduction

While trained evaluation metrics for machine translation (MT) tend to have a high correlation with human judgments (Freitag et al., 2022b), they remain black boxes, sometimes behaving in unexpected ways (Amrhein and Sennrich, 2022; Rei et al., 2023). This calls into question whether a metric's utility can be measured solely by its correlation with human judgments.

In this paper, we intentionally provide MT metrics with *machine-translated reference translations*, as opposed to human-created references, and investigate how this factor influences the behavior of a metric. In MT evaluation research, the human translators who create reference translations are usually asked to produce them from scratch, in order to avoid references that are machine-translated or post-edited (Kocmi et al., 2022). Nevertheless, traces of MT have been detected in some reference sets (Kloudová et al., 2021; Akhbardeh et al., 2021; Kocmi et al., 2022). It is therefore important to understand how metrics behave under such references.

In our experiments, we use a surrogate for real post-edited references in the form of error-free output by various systems from the WMT 2021 news translation task (Akhbardeh et al., 2021). Our results show that there is a stark difference between trained and non-trained metrics: While trained metrics maintain most of their accuracy when provided with such MT-derived references, non-trained metrics exhibit a substantial drop in accuracy.

To corroborate this observation, we perform a controlled experiment involving Prism (Thompson and Post, 2020), a metric that is based on a multilingual MT system. The original version of Prism can be considered non-trained, since it learns from parallel sentences without human judgments. We then fine-tune Prism on a dataset of human judgments, using a bidirectional pairwise ranking approach.

As expected, the segment-level correlation of Prism increases during fine-tuning, indicating that the metric learns to better predict human judgments (Figure 1). Moreover, we find that fine-tuning narrows the gap in performance between human-created and machine-translated references. Our experiment thus indicates that training a metric on human evaluation data can influence its behavior in a way that is not captured by global correlation with human judgments. Code to reproduce our

[*]Work done during an internship at Amazon.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 983–995
December 6–7, 2023.

findings will be made available.[1]

To summarize, the paper makes the following contributions:

- We propose a metric evaluation setup that intentionally uses machine-translated references, and demonstrate that non-trained metrics perform poorly in this setup.
- We present an approach for fine-tuning Prism on human judgments that significantly improves segment-level correlation on unseen test data.
- We show that fine-tuning Prism on human judgments makes it more robust to the use of machine-translated references.

2 Background

2.1 Reference-based Evaluation

Automatic evaluation of MT is often performed by comparing the system output with one or more reference translations, using an evaluation metric. Evaluation metrics can be roughly divided into *trained* and *non-trained* metrics. Trained metrics receive supervision from human judgments of past machine translations. For example, Sellam et al. (2020) and Rei et al. (2020; 'COMET') fine-tuned a pre-trained sentence encoder on such human judgments, using regression or ranking objectives.

Non-trained metrics, on the other hand, rely on a heuristic to make the comparison. Metrics such as BLEU (Papineni et al., 2002) and chrF (Popović, 2015) are based on the overlap of words or characters between the system output and the reference. Thompson and Post (2020) use the perplexity of a neural sequence-to-sequence model, called Prism, that has been trained on multilingual MT. Systematic comparisons of evaluation metrics (Freitag et al., 2022b) have shown that trained metrics tend to correlate better with human judgments than non-trained metrics do, especially if the latter are based on overlap heuristics.

2.2 Quality of Reference Translations

The reliability of reference-based evaluation metrics also depends on the quality of the references they are provided with (Freitag et al., 2021b). A notorious source of noise in references is *translationese*, which is characterized by monotonicity with respect to the source sequence and a high n-gram overlap with system translations (Freitag et al., 2020). Freitag et al. (2020) have shown that translationese references cause BLEU scores to be higher, and the scores are dominated by matches of common, unspecific n-grams. They find that BLEU scores under non-translationese references tend to be lower, but more precise.

Agarwal et al. (2023) observed that post-edited references for spoken language translation seem to inflate BLEU scores, but not the scores of COMET. However, the relationship between metric training and the quality of reference translations has not been studied in detail. In this paper, we hypothesize that robustness to machine-translated references may partially explain why trained metrics are more accurate in practice.

3 Experimental Setup

3.1 Measuring Global Correlation

For measuring the overall correlation of a metric to human judgments, we follow the WMT 2021 metrics task (Freitag et al., 2021b) and use MQM annotations of submissions to the 2021 WMT news translation task (Akhbardeh et al., 2021). The evaluation data cover two domains, news and TED talks. Table A5 reports statistics for these data.

We closely replicate the methodology of the WMT 2021 metrics task. On the segment level, we report Kendall's tau coefficient across all segments and systems; on the system level, we report *pairwise accuracy* (Kocmi et al., 2021), i.e., the ratio of system pairs that a metric ranks in the same order as human annotators have. Following the shared task, we only consider system translations and exclude human translations from the evaluation. We then perform *perm-both* hypothesis tests (Deutsch et al., 2021) to validate metrics comparisons at $\alpha = 0.05$.

3.2 Measuring the Effect of Machine-translated References

In the context of our analysis, we use error-free system translations from the WMT 2021 news translation task as a surrogate for real post-edited references. Specifically, we randomly select system translations that have been annotated according to the MQM standard and in which no annotator has marked an error. This approach allows us to simulate a post-editing process without the cost and noise incurred by actual post-editing.

[1] https://github.com/amazon-science/prism-finetuned

Source sequence (English)
Face masks are mandatory across the state of California, even in fresh air.

Human-created reference (German)
Gesichtsmasken sind im ganzen Bundesstaat Kalifornien vorgeschrieben, auch im Freien.

Machine-translated reference (German)
Gesichtsmasken sind im gesamten Bundesstaat Kalifornien Pflicht, auch an der frischen Luft.

Figure 2: Example of a machine-translated reference compared to the standard reference created by a human translator. The machine-translated reference is more literal (*an der frischen Luft* 'in fresh air').

Figure 2 and Appendix F juxtapose some examples of error-free system translations and the standard, human-created reference translations. The former tend to be more literal and more aligned to the source, both in terms of syntax and content.

It should be noted that when we evaluate a metric in this analysis, we draw from the same set of systems and human annotations as we do for extracting the references. We take care to properly separate the system translations used as a reference from those that are evaluated based on that reference.

To calculate segment-level correlation, we sample a random error-free translation from an unrelated system, for each system output.[2] To calculate system-level pairwise accuracy, we use different sets of references depending on the pair of systems that is compared. Figure 3 shows that our approach is comparable to cross-validation. For every pair of systems that we consider when calculating the pairwise accuracy of a metric, we select one reference translation from an unrelated system, independently per segment. As a consequence, we use slightly different reference sets for ranking different pairs of systems.

We then compare the accuracy of a metric when provided with the machine-translated references to its accuracy when using the standard references. To ensure comparability, we skip all the segments where no machine-translated reference is available (which is either because the segment has not been part of the annotation study or because annotators have found an error in every system translation). The metric accuracies for both ref_{std} and ref_{mt} are

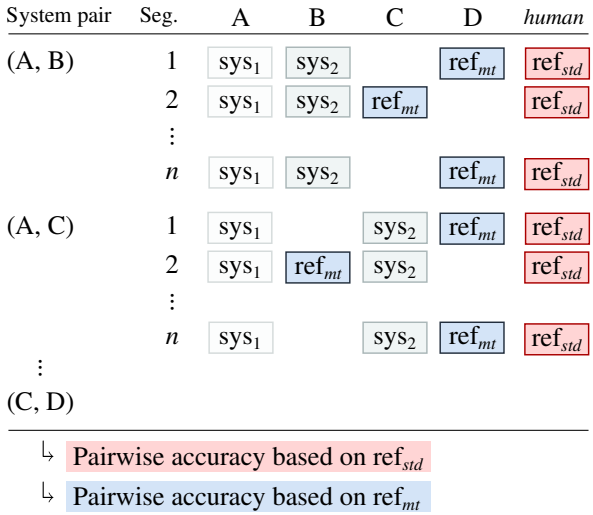


Figure 3: To measure the effect of machine-translated references, we use error-free output from other, unrelated MT systems as references. For example, when comparing system A to system B, we use a translation from either system C, D, etc. as a reference for each segment.

thus calculated based on a subset of the segments used to calculate global correlation. Table A5 shows that only for one language pair a substantial number of segments need to be skipped (Chinese–English news). For the other language pairs, between 0% and 4.5% of the segments are skipped.

4 Fine-tuning the Prism Metric

Prism (Thompson and Post, 2020) is a reference-based evaluation metric that relies on the paraphrasing probability between a system translation and a reference. The probability is estimated by a multilingual NMT model as a zero-shot translation direction. The model is expected to prefer mere copies of the source sequence to more creative paraphrases, which is especially useful for reference-based evaluation.

The NMT model uses the reference as a source

[2] Segment-level correlation is calculated jointly across all segments and systems, and as a consequence, using different references to evaluate the translations of different systems adds some noise to the correlation. However, we expect that the correlation is dominated by the segment axis and not by the system axis. Our findings on the segment level are consistent with our findings on the system level.

sequence x and the system translation as a hypothesis y, or vice versa. The segment-level score S is then calculated from token-level log-probabilities:[3]

$$S(y|x) = \frac{1}{|y|}\sum_{t=1}^{|y|}\log p(y_t|y_{i<t}, x).$$

By default, Prism uses the average of both paraphrasing directions:

$$\text{Prism}(\text{sys}, \text{ref}) = \frac{1}{2}S(\text{sys}|\text{ref}) + \frac{1}{2}S(\text{ref}|\text{sys}).$$

An overall score for a system can then be calculated as an average over a collection of segments.

4.1 Training Objective

In order to fine-tune Prism, we combine a standard cross-entropy objective and a bidirectional pairwise ranking objective.

For the *cross-entropy objective*, we use the source sequence (src) and the reference translation (ref) of the training examples to continue the cross-entropy training:

$$L_{\text{src}\to\text{ref}} = -S(\text{ref}|\text{src}).$$

Our goal in using this objective is to familiarize Prism with the segments to which the human judgments refer, and to prevent catastrophic forgetting during the fine-tuning stage.

In addition, we propose a *bidirectional pairwise ranking objective*. In the forward direction, we train Prism to correctly rank two system translations (sys^+ and sys^-), conditioned on the reference (*forward ranking*):

$$L_{\text{ref}\to\text{sys}} = \max\{0, \epsilon - S(\text{sys}^+|\text{ref}) + S(\text{sys}^-|\text{ref})\},$$

where ϵ is a margin value. We add a second ranking loss for the reverse paraphrasing direction, i.e., for reconstructing the reference from either of the system translations (*backward ranking*):

$$L_{\text{sys}\to\text{ref}} = \max\{0, \epsilon - S(\text{ref}|\text{sys}^+) + S(\text{ref}|\text{sys}^-)\}.$$

The complete fine-tuning objective is:

$$L = \alpha L_{\text{src}\to\text{ref}} + (\frac{1}{2}L_{\text{ref}\to\text{sys}} + \frac{1}{2}L_{\text{sys}\to\text{ref}}),$$

where α is a scalar to balance the two terms.

Figure 4 is a schematic illustration of the objectives for pre-training, fine-tuning, and inference.

[3]This score is called H in the original definition. We use S instead, to avoid confusion with cross-entropy (which is $-S$).

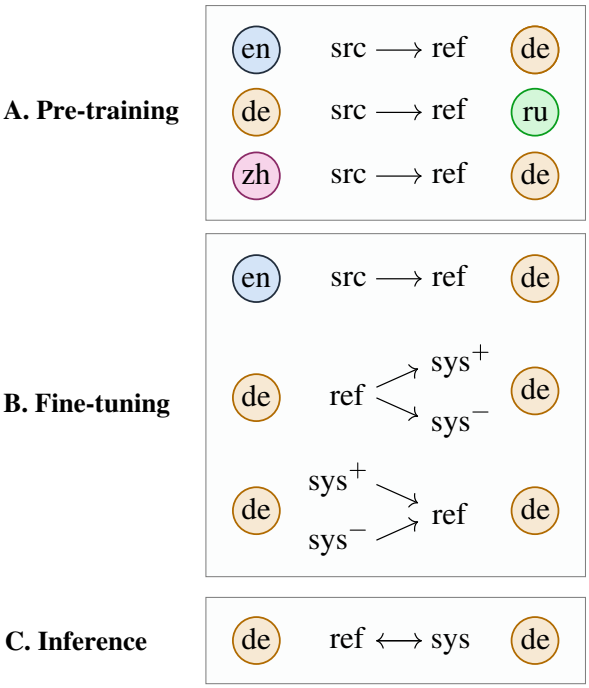


Figure 4: Schematic illustration of the sequences used for pre-training, fine-tuning, and applying the Prism model to MT evaluation. Prism has been **(A)** pre-trained on multilingual translation to and from 39 languages as described by Thompson and Post (2020); inference **(C)** makes use of the zero-shot paraphrasing capability acquired by the model during pre-training. We add a fine-tuning stage **(B)** with data derived from human evaluations of MT. In this illustration, Prism is fine-tuned on English–German examples.

4.2 Training Data

For fine-tuning Prism, we use human judgments of submissions to the 2020 WMT news translation tasks (Barrault et al., 2020), collected by Freitag et al. (2021a).[4] These annotations are based on the Multidimensional Quality Metrics (MQM) framework (Lommel et al., 2014) and have been shown to correlate better with automatic metrics than previous direct assessments, especially when the evaluation concerns high-quality translations (Freitag et al., 2021a,b). Specifically, we train Prism on human judgments for English–German and Chinese–English translations of news. We train a single model jointly on both language pairs.

To use the human judgments for training on pairwise ranking, the direct MQM assessments need to be converted into relative rankings of translation pairs. In previous work, direct (non-MQM) assess-

[4]Submission data are available at https://github.com/google-research/mt-metrics-eval and the MQM annotations are available at https://github.com/google/wmt-mqm-human-evaluation

ments have been normalized and aggregated across annotators before being compared (Ma et al., 2019). Since MQM ratings are known to have low inter-annotator agreement on the segment level (Freitag et al., 2021b), we opt for intra-annotator pairing instead. Specifically, we only pair translations that have been rated by the same annotator, and we do not compare MQM scores across annotators. Relative rankings are created independently for each annotator and then concatenated. Furthermore, we only pair translations that have a score difference greater than 0.1, which would correspond to a minor fluency or punctuation error. Taken together, these criteria should ensure there is a noticeable difference between the quality of two system translations sys^+ and sys^- in the eyes of at least one annotator. We hold out 5000 relative rankings from the resulting training data as a validation set and use it to select hyperparameters. Detailed statistics for the training data are provided in Table A4.

4.3 Implementation Details

The fine-tuning was implemented in Fairseq (Ott et al., 2019). We start with the original `Prism39` model released by Thompson and Post (2020).[5] We then fine-tune the model for a single pass over the training data, using Adam. The initial learning rate is set to 1e-4 without any warm-up steps. We use half-precision training and an effective batch size of 360k tokens. Other settings match the pre-training setup of Prism.

We set the margin hyperparameter ϵ to 0.1, and the cross-entropy weight α to 0.1 as well. The hyperparameters have been selected based on segment-level correlation on the validation set. Since we jointly train on two language pairs, we iterate over batches for each language pair in a round-robin fashion, upsampling the smaller language pair. Fine-tuning takes about one hour on a `p3.8xlarge` AWS instance, which has 4 Tesla V100 GPUs with 16 GB of memory.

5 Results

Effect of fine-tuning Prism Table 1 shows that fine-tuning Prism has the intended effect: *Fine-tuning Prism on human judgments of machine translations significantly improves correlation with human judgments on an unseen test set.* The effect of fine-tuning is especially pronounced for the English–German and Chinese–English language pairs, since the metric was fine-tuned on those pairs. Interestingly, we also observe positive cross-lingual transfer to the English–Russian language pair, which was not seen during fine-tuning. Table 2 shows that the positive effect of fine-tuning extends to the TED Talks domain, even though the metric was not fine-tuned on this domain.

	EN–DE	EN–RU	ZH–EN
Prism	19.3	22.4	28.8
Prism+FT	**25.3**	**23.7**	**31.5**

Table 1: In-domain accuracy of Prism on WMT 2021 news translation submissions. We report segment-level Kendall's tau correlation to human judgments. Bold font denotes that the improvement achieved through fine-tuning is significant with $\alpha = 0.05$. Note that Prism+FT has not been fine-tuned on the EN–RU language pair.

	EN–DE	EN–RU	ZH–EN
Prism	24.2	21.9	19.6
Prism+FT	**26.9**	22.3	**21.9**

Table 2: Out-of-domain accuracy of Prism on WMT 2021 system translations of TED talks in terms of segment-level Kendall's tau. Bold indicates that the improvement is significant with $\alpha = 0.05$.

Effect of using machine-translated references Table 3 reports the segment-level correlation of different metrics when using either standard references or machine-translated references. Note that the values for Prism slightly differ from Tables 1 and 2 because this analysis is based on a subset of the segments. *We find that the correlation of metrics to human judgments tends to decrease under machine-translated references.* For the Chinese–English dataset the relative decline is smaller than average, but is still noticeable for most metrics.

In Table 3, when comparing the non-trained metrics (above the horizontal line) to the trained metrics (below the line), we observe that the decline in correlation is smaller for the trained metrics. An especially interesting comparison is between Prism and Prism+FT, given that the two metrics differ only in the training data. *Prism+FT is consistently more robust to machine-translated references than Prism, indicating that the metric learns to cope with such references during the fine-tuning stage.*

With respect to system-level pairwise accuracy (Table 4), we observe a similar trend.

[5] https://data.statmt.org/prism/

	EN−DE		EN−RU		ZH−EN		Average	
	ref_{std}	ref_{mt}	ref_{std}	ref_{mt}	ref_{std}	ref_{mt}	ref_{std}	ref_{mt}
BLEU	8.4	7.0 **(-16.7%)**	12.1	11.8 **(-2.5%)**	15.2	14.8 **(-2.6%)**	11.9	11.2 **(-5.9%)**
chrF	11.1	8.3 **(-25.2%)**	19.3	13.8 **(-28.5%)**	16.7	15.7 **(-6.0%)**	15.7	12.6 **(-19.7%)**
Prism	18.9	18.2 **(-3.7%)**	22.4	20.6 **(-8.0%)**	24.2	23.5 **(-2.9%)**	21.8	20.8 **(-4.9%)**
Prism+FT	24.9	24.4 **(-2.0%)**	23.7	22.3 **(-5.9%)**	26.6	26.8 **(0.8%)**	25.1	24.5 **(-2.3%)**
COMET	25.1	24.6 **(-2.0%)**	27.6	25.4 **(-8.0%)**	32.1	32.1 **(0.0%)**	28.3	27.4 **(-3.2%)**

Table 3: Segment-level correlation of MT metrics when provided with the standard references (ref_{std}) of the WMT21 metrics news subtask (Freitag et al., 2021b), and with machine-translated references (ref_{mt}). The percentages denote the relative change in correlation when falling back to machine-translated references. The trained metrics, Prism+FT and COMET (`wmt21-comet-mqm`), have a more favorable relative change than the non-trained metrics, which indicates higher robustness to machine-translated references.

	EN−DE		EN−RU		ZH−EN		Average	
	ref_{std}	ref_{mt}	ref_{std}	ref_{mt}	ref_{std}	ref_{mt}	ref_{std}	ref_{mt}
BLEU	89.7	74.4 **(-17.1%)**	70.3	58.2 **(-17.2%)**	61.5	61.5 **(0.0%)**	73.8	64.7 **(-12.4%)**
chrF	87.2	71.8 **(-17.7%)**	74.7	56.0 **(-25.0%)**	60.3	56.4 **(-6.5%)**	74.1	61.4 **(-17.1%)**
Prism	85.9	73.1 **(-14.9%)**	83.5	62.6 **(-25.0%)**	61.5	56.4 **(-8.3%)**	77.0	64.0 **(-16.8%)**
Prism+FT	89.7	80.8 **(-9.9%)**	80.2	61.5 **(-23.3%)**	61.5	61.5 **(0.0%)**	77.1	67.9 **(-11.9%)**
COMET	79.5	84.6 **(6.4%)**	68.1	65.9 **(-3.2%)**	60.3	55.1 **(-8.6%)**	69.3	68.5 **(-1.1%)**

Table 4: System-level pairwise accuracy of MT metrics when provided with the standard references of the WMT21 metrics news subtask (Freitag et al., 2021b), and with machine-translated references. Again, the trained metrics, Prism+FT and COMET (`wmt21-comet-mqm`), tend to be more robust to machine-translated references.

Prism+FT does not show significantly higher pairwise accuracy than Prism when using standard references, which is explained by the high statistical variance of the pairwise accuracy metric. But again, Prism+FT appears more robust to machine-translated references than Prism. Finally, Appendix B reports results for the TED talks domain, where the same patterns can be observed.

Ablation Study We perform an ablation study to measure the influence to the three terms in the Prism fine-tuning objective. Appendix A shows that removing either of the three terms decreases segment-level correlation. The ablation shows that the cross-entropy objective has the additional effect of stabilizing the model: Without cross-entropy, the average probability scores output by Prism shift from 0.47 to 0.35 after a single epoch of fine-tuning, and the BLEU achieved by the Prism translation model on an unseen test set clearly declines.

6 Related Work

Machine translations as references Popovic et al. (2016) first investigated the potential of using post-edited machine translations as references, finding that post-edited translations stemming from high-quality systems are better references than those from low-quality systems. Toral (2019) argued that post-edited machine translations can be seen as an exacerbated form of translationese (*post-editese*). Combined with the finding of Freitag et al. (2020) that translationese references are less favorable than intentionally paraphrased references, this suggests that machine translations, even if post-edited, are a challenge for MT evaluation.

Albrecht and Hwa (2007) propose to train an evaluation metric using non-annotated translations of other systems as *pseudo-references*. They hypothesize that a metric can learn to detect and to constructively utilize any errors in these references. Yoshimura et al. (2019) instead use a paraphrase identifier to filter pseudo-references based on their paraphrastic similarity to a human-created reference. Finally, minimum Bayes risk decoding (Kumar and Byrne, 2004) employs pseudo-references for generating translations, and has been shown to depend on robust metrics as well (Freitag et al., 2022a; Amrhein and Sennrich, 2022).

Training a sequence-to-sequence model on pairwise ranking Pairwise ranking has commonly been used to train SVM (Ye et al., 2007; Duh, 2008; Stanojević and Sima'an, 2014) and neural network encoders (Guzmán et al., 2015; Dušek et al., 2019). A more recent approach has been to fine-tune pre-trained sentence encoders so that the embedding similarities of two hypotheses and the reference and/or source are optimized for pairwise ranking (Rei et al., 2020; Zhang and van Genabith, 2020), in which case the max-margin loss reduces to a triplet margin loss (Schroff et al., 2015). In this paper, we do not rely on the similarity of sentence embeddings but use the perplexity of a sequence-to-sequence model as a metric.

Since we optimize perplexity given positive and negative examples, our fine-tuning approach becomes very similar to contrastive learning for NMT. Typical applications of contrastive learning try to eliminate specific translation error types by creating perturbed versions of the training references (Yang et al., 2019; Hwang et al., 2021). A similar objective has been used for discriminative re-ranking of translation candidates (Shen et al., 2004; Yu et al., 2020). In this paper, however, the goal is not to improve translation output but to train an evaluation metric on human judgments.

7 Conclusion

We have shown that metrics without supervision by human judgments, such as BLEU and chrF, tend to be inaccurate under machine-translated references, while trained metrics are more robust. In order to methodically examine this phenomenon, we have trained the Prism evaluation metric on a dataset of human judgments. Our experiments show that fine-tuning improves the segment-level accuracy of Prism on an unseen test set across multiple language pairs and domains, and clearly increases its robustness to machine-translated references.

One conclusion to draw from our findings is that post-edited references likely diminish the accuracy of reference-based metrics and should be avoided. A second conclusion is that if it cannot be ruled out that references originate from MT, as is often the case in practice, trained metrics are to be preferred. Fine-tuning a metric such as Prism on reference-based evaluation can thus be seen as a technique to let the metric make the best out of reference translations in the wild.

Limitations

Our study is mainly limited by the data we use for fine-tuning and evaluating Prism. The experiments are based on three language pairs only. Automatic MT evaluation is relevant for many more language pairs and language families, including and maybe especially so for low-resource settings.

Secondly, it should be mentioned that the machine translations we use in our analysis have been generated by systems based on a similar technology. Almost all of the systems seem to use the Transformer architecture, and they have all been trained on similar data (Akhbardeh et al., 2021). It is possible that our findings do not generalize to the evaluation of other varieties of MT, such as rule-based systems, or to reference-based evaluation metrics that use large language models (Kocmi and Federmann, 2023).

Acknowledgements

We thank Bill Byrne, Felix Hieber, Brian Thompson and Ke Tran for comments on an earlier stage of this project. JV and RS acknowledge funding by the Swiss National Science Foundation (project MUTAMUR; no. 176727).

References

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Kevin Duh, Yannick Estève, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutai Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Kr. Ojha, John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 2023. Findings of the IWSLT 2023 evaluation campaign. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa,

Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Joshua Albrecht and Rebecca Hwa. 2007. Regression for sentence-level MT evaluation with pseudo references. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 296–303, Prague, Czech Republic. Association for Computational Linguistics.

Chantal Amrhein and Rico Sennrich. 2022. Identifying weaknesses in machine translation metrics through minimum Bayes risk decoding: A case study for COMET. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1125–1141, Online only. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146.

Kevin Duh. 2008. Ranking vs. regression in machine translation evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 191–194, Columbus, Ohio. Association for Computational Linguistics.

Ondřej Dušek, Karin Sevegnani, Ioannis Konstas, and Verena Rieser. 2019. Automatic quality estimation for natural language generation: Ranting (jointly rating and ranking). In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 369–376, Tokyo, Japan. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022a. High Quality Rather than High Model Probability: Minimum Bayes Risk Decoding with Neural Metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022b. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2015. Pairwise neural machine translation evaluation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 805–814, Beijing, China. Association for Computational Linguistics.

Yongkeun Hwang, Hyeongu Yun, and Kyomin Jung. 2021. Contrastive learning for context-aware neural machine translation using coreference information. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1135–1144, Online. Association for Computational Linguistics.

Věra Kloudová, Ondřej Bojar, and Martin Popel. 2021. Detecting post-edited references and their effect on human evaluation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 114–119, Online. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp

Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popovic, Mihael Arčan, and Arle Lommel. 2016. Potential and limits of using post-edits as reference translations for MT evaluation. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 218–229.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, Marcos Treviso, Luisa Coheur, Alon Lavie, and André Martins. 2023. The inside story: Towards better understanding of machine translation neural evaluation metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1089–1105, Toronto, Canada. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 815–823. IEEE Computer Society.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Libin Shen, Anoop Sarkar, and Franz Josef Och. 2004. Discriminative reranking for machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 177–184, Boston, Massachusetts, USA. Association for Computational Linguistics.

Miloš Stanojević and Khalil Sima'an. 2014. Fitting sentence level translation evaluation with many dense features. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 202–206, Doha, Qatar. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Antonio Toral. 2019. Post-editese: an exacerbated translationese. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 273–281, Dublin, Ireland. European Association for Machine Translation.

Zonghan Yang, Yong Cheng, Yang Liu, and Maosong Sun. 2019. Reducing word omission errors in neural machine translation: A contrastive learning approach. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6191–6196, Florence, Italy. Association for Computational Linguistics.

Yang Ye, Ming Zhou, and Chin-Yew Lin. 2007. Sentence level machine translation evaluation as a ranking. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 240–247, Prague, Czech Republic. Association for Computational Linguistics.

Ryoma Yoshimura, Hiroki Shimanaka, Yukio Matsumura, Hayahide Yamagishi, and Mamoru Komachi. 2019. Filtering pseudo-references by paraphrasing for automatic evaluation of machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 521–525, Florence, Italy. Association for Computational Linguistics.

Lei Yu, Laurent Sartran, Po-Sen Huang, Wojciech Stokowiec, Domenic Donato, Srivatsan Srinivasan, Alek Andreev, Wang Ling, Sona Mokra, Agustin Dal Lago, Yotam Doron, Susannah Young, Phil Blunsom, and Chris Dyer. 2020. The DeepMind Chinese–English document translation system at WMT2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 326–337, Online. Association for Computational Linguistics.

Jingyi Zhang and Josef van Genabith. 2020. Translation quality estimation by jointly learning to score and rank. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2592–2598, Online. Association for Computational Linguistics.

A Ablation Study

Variant	Segment-level Kendall's tau	Pairwise accuracy	Magnitude of scores	BLEU (newstest21) EN−DE	BLEU (newstest21) ZH−EN
Prism (no fine-tuning)	23.5	78.7	0.47	25.6	18.7
Prism+FT	26.8	76.7	0.37	23.0	21.0
– without cross-entropy	26.6	74.6	0.35	10.2	9.6
– without forward ranking	26.0	79.2	0.40	21.9	20.1
– without backward ranking	25.6	77.7	0.39	21.1	20.3

Table A1: Ablation study for the proposed fine-tuning objective, based on the in-domain meta-evaluation setting (WMT 2021 news translations). In every row we remove one aspect of the fine-tuning setup. Meta-metrics are averaged across three language pairs. *Magnitude of scores* refers to the average segment-level scores predicted by the Prism model, converted to probability space via 2^x.

B Evaluation on TED Talks

	EN−DE		EN−RU		ZH−EN		Average	
	ref_{std}	ref_{mt}	ref_{std}	ref_{mt}	ref_{std}	ref_{mt}	ref_{std}	ref_{mt}
BLEU	13.4	7.1 **(-47.0%)**	16.0	12.8 **(-20.0%)**	11.0	9.1 **(-17.3%)**	13.5	9.7 **(-28.2%)**
chrF	14.3	7.9 **(-44.8%)**	18.9	12.8 **(-32.3%)**	11.4	9.0 **(-21.1%)**	14.9	9.9 **(-33.4%)**
Prism	23.6	17.7 **(-25.0%)**	22.0	17.5 **(-20.5%)**	18.0	15.9 **(-11.7%)**	21.2	17.0 **(-19.7%)**
Prism+FT	26.4	24.2 **(-8.3%)**	22.2	21.6 **(-2.7%)**	20.2	19.4 **(-4.0%)**	22.9	21.7 **(-5.2%)**
COMET	27.3	24.6 **(-9.9%)**	25.8	23.2 **(-10.1%)**	20.8	20.7 **(-0.5%)**	24.6	22.8 **(-7.3%)**

Table A2: Segment-level correlation of MT metrics when provided with the standard references and with machine-translated references. The percentages denote the relative change in correlation when falling back to machine-translated references.

	EN−DE		EN−RU		ZH−EN		Average	
	ref_{std}	ref_{mt}	ref_{std}	ref_{mt}	ref_{std}	ref_{mt}	ref_{std}	ref_{mt}
BLEU	66.7	35.9 **(-46.2%)**	83.5	58.2 **(-30.3%)**	64.1	65.4 **(2.0%)**	71.4	53.2 **(-25.6%)**
chrF	65.4	46.2 **(-29.4%)**	85.7	53.8 **(-37.2%)**	61.5	66.7 **(8.5%)**	70.9	55.6 **(-21.6%)**
Prism	69.2	44.9 **(-35.1%)**	82.4	48.4 **(-41.3%)**	67.9	66.7 **(-1.8%)**	73.2	53.3 **(-27.1%)**
Prism+FT	66.7	51.3 **(-23.1%)**	81.3	61.5 **(-24.4%)**	62.8	70.5 **(12.3%)**	70.3	61.1 **(-13.0%)**
COMET	84.6	53.8 **(-36.4%)**	78.0	74.7 **(-4.2%)**	67.9	75.6 **(11.3%)**	76.8	68.0 **(-11.5%)**

Table A3: System-level pairwise accuracy of MT metrics when provided with the standard references and with machine-translated references.

C Training Data Statistics

Language pair	**EN–DE**	**ZH–EN**
Number of systems (including sets of human translations)	10	10
Number of annotated segments	1 418	2 000
– used for relative rankings	1 411	1 985
Number of annotated system translations	14 110	19 994
– used for relative rankings	14 110	19 850
Number of relative rankings	126 217	164 137
– training split	121 217	159 137
– validation split	5 000	5 000

Table A4: Statistics for the WMT 2020 MQM ratings (Freitag et al., 2021a) and for the relative rankings that we derive using an intra-annotator pairing approach.

D Meta-Evaluation Data Statistics

	News			TED Talks		
	EN–DE	EN–RU	ZH–EN	EN–DE	EN–RU	ZH–EN
Number of systems (without human)	13	14	13	13	14	13
Number of MQM-annotated segments	527	527	650	529	512	529
Number of segments with machine-translated reference (on average across system pairs)	518	527	461	517	511	505

Table A5: Statistics for the WMT 2021 MQM ratings (Freitag et al., 2021b) we use for evaluating the metrics.

E Model Hyperparameters

Model	N	d_{model}	d_{ffn}	h	Parameters	Vocabulary size
Prism (Thompson and Post, 2020)	16	1280	12288	20	745M	64k
`wmt21-comet-mqm` (Rei et al., 2021)	24	1024	4096	16	581M	250k

Table A6: Hyperparameters of the Transformer-based metrics.

F Additional Examples of Human-created and Machine-translated References

English–German News Example

Source sequence:
Face masks are mandatory across the state of California, even in fresh air.

Standard reference:
Gesichtsmasken sind im ganzen Bundesstaat Kalifornien vorgeschrieben, auch im Freien.

Randomly sampled error-free system translation (Nemo):
Gesichtsmasken sind im gesamten Bundesstaat Kalifornien Pflicht, auch an der frischen Luft.

Chinese–English News Example

Source sequence:
他已承认，是自己在教堂里点火。

Standard reference:
The parish volunteer has admitted that he had started the fire in the church.

Randomly sampled error-free system translation (metricsystem5):
He has admitted that it was himself who set the fire in the church.

English–German TED Talks Example

Source sequence:
Today I'd like to show you the future of the way we make things.

Standard reference:
Ich möchte Ihnen heute zeigen, wie wir in Zukunft Dinge herstellen werden.

Randomly sampled error-free system translation (Online-W):
Heute möchte ich Ihnen die Zukunft der Art und Weise zeigen, wie wir Dinge herstellen.

Chinese–English TED Talks Example

Source sequence:
今天我想向各位展示未来我们制作东西的方式。

Standard reference:
Today I'd like to show you the ways we make things in the future.

Randomly sampled error-free system translation (metricsystem1):
Today I want to show you how we will make things in the future.

Training and Meta-Evaluating Machine Translation Evaluation Metrics at the Paragraph Level

Daniel Deutsch, Juraj Juraska, Mara Finkelstein, and Markus Freitag
Google
{dandeutsch,jjuraska,marafin,freitag}@google.com

Abstract

As research on machine translation moves to translating text beyond the sentence level, it remains unclear how effective automatic evaluation metrics are at scoring longer translations. In this work, we first propose a method for creating paragraph-level data for training and meta-evaluating metrics from existing sentence-level data. Then, we use these new datasets to benchmark existing sentence-level metrics as well as train learned metrics at the paragraph level. Interestingly, our experimental results demonstrate that using sentence-level metrics to score entire paragraphs is equally as effective as using a metric designed to work at the paragraph level. We speculate this result can be attributed to properties of the task of reference-based evaluation as well as limitations of our datasets with respect to capturing all types of phenomena that occur in paragraph-level translations.

1 Introduction

Automatic evaluation metrics have always been a critical component to the progress of research on machine translation (MT). As the field of MT moves beyond translating individual sentences to translating full paragraphs, book chapters, or documents (Tu et al., 2018; Sun et al., 2022; Thai et al., 2022; Jiang et al., 2023; Post and Junczys-Dowmunt, 2023), automatic metrics need to be designed to work on these longer texts.

Currently, how well automatic metrics agree with human judgments of paragraph translation quality is an open question.[1] Few studies have meta-evaluated metrics on longer texts, and those that have are focused on the literary domain and are limited in the size of the evaluation dataset (Jiang et al., 2022; Thai et al., 2022; Karpinska and Iyyer, 2023). In this work, we investigate training and meta-evaluating metrics for scoring paragraph translations using the benchmark Workshop on Machine Translation (WMT) datasets that are widely used for metric development (Freitag et al., 2022).

Due to the scarcity of human ratings of paragraph translations, we propose a method to create paragraph-level training and meta-evaluation datasets from the existing WMT sentence-level datasets (§3). Although these ratings are typically only used at the sentence level, they were collected on contiguous paragraphs and performed with document context, so they can be used as paragraph-level datasets. We repurpose these datasets to benchmark existing sentence-level metrics as well as train new paragraph-level metrics for scoring paragraph translations (§4).

Our experimental results are somewhat surprising. We find that there appears to be little evidence that training on paragraph-level data is beneficial—at least given the limitations of our experimental setup. Using metrics trained on sentence-level data only to directly score full paragraphs achieves comparable agreement to human ratings as metrics trained on paragraph-level data (§6.1). Sentence-level metrics appear to generalize well to inputs much longer than they were trained on (§6.2).

We hypothesize these observations can be explained by the nature of evaluating translations and characteristics of our paragraph-level dataset (§7). We speculate that long range dependencies—which paragraph-level metrics can model but sentence-level likely do not—may not be too important for achieving high agreement with human ratings. Further, due to the fact that our training and evaluation datasets assume a sentence alignment between the reference and hypothesis paragraphs, certain translation phenomena that sentence-level metrics may struggle to handle, like sentence or information reordering, are not well represented in the dataset,

[1] Translation beyond the sentence level is often referred to as document-level MT. However, there is no clear definition for the term "document." We use "paragraph" in this work because we feel it most accurately describes the length of text in our datasets. See §2 for more details on this.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 996–1013
December 6–7, 2023.

limiting our ability to show the benefits of training on paragraph-level ratings.

The contributions of our work include (1) a method for constructing paragraph-level training and meta-evaluation datasets from sentence-level ratings, (2) an experimental study that demonstrates the comparable performance of sentence- and paragraph-level metrics, and (3) an analysis that aims to provide an explanation for our experimental observations.

2 Terminology

Throughout this paper, we use terms like segment, sentence, paragraph, and document to refer to different lengths of text. To the best of our knowledge, there are no agreed upon definitions for these terms in the MT literature, so here we define how they are used for the rest of the paper.

We refer to the input text to an MT system or evaluation metric as a *segment*, irrespective of its length. Traditionally, segments in MT have been roughly equivalent to one sentence, although sometimes they can be short phrases or even longer than a single sentence. Regardless, we use *sentence* to refer to this unit of text since it accurately describes the most common text length that is widely used in MT.

Our work investigates evaluating *paragraphs* of text, which we define to be multi-sentence segments. We do not require that the paragraphs used in this work obey the traditional definition of a paragraph (i.e., a unit of text separated by a newline character). We refrain from calling this unit of text a *document*—which we consider to be all of the possible input text—since each document can be broken down into multiple paragraphs and the term paragraph more accurately describes the length of text we use.

3 Paragraph-Level Datasets

The two main sources for training and meta-evaluating MT metrics are the direct assessment (DA) and Multi-dimensional Quality Metrics (MQM; Lommel et al., 2014; Freitag et al., 2021a) datasets that the Workshop on Machine Translation (WMT) has collected as part of the yearly metrics shared task (Freitag et al., 2022). The DA ratings were done by a mixture of expert and non-expert raters (depending on whether the translation direction is into or out of English) who assigned a quality score in the range 0-100 to translated sentences.

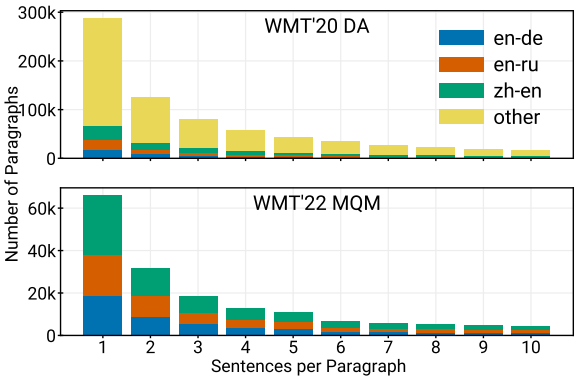


Figure 1: The number of contiguous paragraphs for the given number of sentences per paragraph where each sentence is rated by the same rater. Actual values are included in Appendix A.

Because of differences in rater behavior, the DA scores are z-normalized per rater.[2] In MQM, expert raters identify error spans in translated sentences and assign each error a category and severity level, which are used to calculate a score for that error. A sentence's MQM score is defined as the sum of the errors' scores.

Training and meta-evaluating metrics at the paragraph level requires a collection of translated paragraphs and paragraph-level quality scores. Luckily, the DA data since 2019 and the MQM data can be considered to be paragraph-level ratings. The ratings were performed on contiguous blocks of sentences that were translated by the same system (e.g., the first k sentences per document are rated for a system). Although the scores were collected at the sentence level, the ratings were done in context, meaning the raters had access to the document context for a sentence, so the scores should reflect paragraph- or document-level phenomena like discourse errors. Therefore, we use the sentence-level DA and MQM data to construct paragraph-level datasets as follows.

For each document translated by a system, we run sliding window of size k sentences from the start to the end. If all k sentences in the window have been rated, those k sentences are concatenated together to become a paragraph instance and the window shifts by k. Otherwise, the sliding window shifts by 1 and the process repeats. To maintain consistency between the sentence scores within a paragraph, we additionally require that every sen-

[2]The methodology for collecting DA ratings has changed throughout the years. See Barrault et al. (2020) for the description in 2020, the most recent year used in this work.

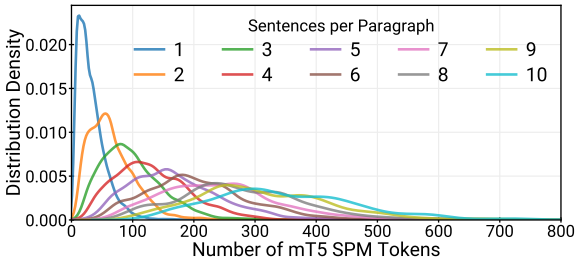


Figure 2: The distribution of paragraph lengths in SPM tokens (i.e., sub-word tokens; Kudo and Richardson, 2018) on the WMT'22 MQM dataset for different numbers of sentences per paragraph. Additional datasets' distributions are included in Appendix A.

tence is scored by the same rater. Then, we define the paragraph-level scores to be the average DA z-score or sum of MQM scores for each sentence in the paragraph.[3] The result is a dataset of rated paragraph translations of k sentences each.

We apply this dataset construction approach to the DA and MQM data for $k = 1, 2, \ldots, 10$ sentences per paragraph. The number of paragraphs is shown in Figure 1 and the distribution of the lengths of the new translated paragraphs is shown in Figure 2. As k increases, the number of paragraphs decreases because there are fewer candidate paragraphs, while the length of the paragraphs increases, roughly by an expected factor of k.

These paragraph-level DA and MQM datasets are used to train and meta-evaluate paragraph-level metrics for the rest of this paper.

4 Paragraph-Level Metrics

We explore two different methods for creating paragraph metrics: directly applying sentence-level metrics to paragraphs (§4.1) and training metrics on paragraph-level data (§4.2).

4.1 Applying Sentence-Level Metrics on Paragraphs

Although automatic metrics that have been used to evaluate sentence-level MT were not explicitly designed to evaluate paragraphs, they can be repurposed to score paragraphs in different ways.

First, the input paragraph can be treated as if it were one long segment and passed to the metric to calculate a score. For metrics that use bag-of-n-grams representations, like BLEU (Papineni et al., 2002), there is no input length limitation. However, some learned metrics, like BLEURT (Sellam et al., 2020), have a maximum possible sequence length due to restrictions related to neural network architectures. Therefore, the length of the input paragraph is restricted in some cases.

Then, if there is assumed to be an alignment between the source, reference, and hypothesis sentences within a paragraph (as is in the case with our datasets), a paragraph score can be calculated by averaging the sentence-level metric's score for each of the k individual sentences. While this sliding window approach more closely aligns how the metrics are being used to how they were designed, we argue this approach is less than ideal because the 1:1 sentence alignment between the source and hypothesis translations will not always exist. However, this approach is useful for understanding and analyzing the behavior of metrics when they are used to score full paragraphs directly.

4.2 Learning Paragraph-Level Metrics

While sentence-level metrics can be repurposed to score paragraphs, the lengths of the input paragraphs are significantly longer than the lengths of individual sentences (compare $k = 1$ to $k > 1$ in Figure 2) and there may be cross-sentence dependencies that are not learned by sentence-level metrics. Therefore, we explore creating a metric specifically for paragraph-level data.

To do so, we train a BLEURT-style regression model on the paragraph-level datasets: The reference and hypothesis paragraphs are tokenized and concatenated together (separated by a special token), then passed as input to a neural network. The network is then trained to predict the hypothesis paragraph's ground-truth quality score. Sections 5.2 and 5.4 contain more information about the model's architecture and implementation details.

It is desirable for the paragraph-level metric to be able to score paragraphs of any length, so we train the metric on paragraphs composed of $k = 1, 2, \ldots, 10$ sentences. Because the number of paragraph instances decreases significantly as k increases (see Figure 1), longer paragraphs will rarely be seen during training. Therefore, we explore two different techniques for weighting training data: one that selects paragraphs uniformly at random

[3] Summing MQM scores was done to generalize an MQM rating for paragraphs since a sentence's MQM score is the total error weight for that sentence. The choice of summing or averaging does not matter for metric meta-evaluation because the correlations are scale invariant.

and one that performs a stratified sample so the training data is composed of an equal number of paragraphs for each value of k.

Next, we describe the experimental setup to evaluate the paragraph-level metrics.

5 Experimental Setup

5.1 Datasets

The paragraph-level datasets used in our experiments are described in Section 3. The WMT'19 (Ma et al., 2019) and '20 (Mathur et al., 2020) paragraph-level DA data is used for training the metrics described in this work, and all metrics are evaluated on the WMT'21 (Freitag et al., 2021b) and WMT'22 (Freitag et al., 2022) paragraph-level MQM data. For both DA and MQM, we use $k = 1, 2, \ldots, 10$ sentences per paragraph. The different paragraph lengths are combined during training but separated for evaluation.

We additionally analyze the behavior of the metrics that we train on judgments collected by Karpinska and Iyyer (2023) on literary translations. Their dataset contains human preference judgments between paragraph translations. The translations come from translation models that translated the input one sentence a time in isolation, one sentence at a time in context, the full paragraph directly, and Google Translate. We evaluate how frequently the metrics agree with the human preference judgments.

5.2 Metrics

Paragraph-Level Metrics. We train two different paragraph-level metrics, one for each of the different weighting techniques, uniform and stratified sampling (see §4.2). We refer to these metrics as PARA-UNIF and PARA-STRAT.

Our metric uses the same architecture as the Metric-X WMT'22 metrics shared task submission (Freitag et al., 2022). The metric builds on the mT5 encoder-decoder language model (Xue et al., 2021), which was originally designed to be a sequence-to-sequence language model. We repurpose the model for our regression task as follows. The inputs to the encoder are the hypothesis and reference translations separated by a special token, and a single dummy token is passed as the first input to the decoder. We arbitrarily selected a reserved vocabulary token, then trained the model so that token's output logit in the first decoding step becomes the score for the input hypothesis translation. This modification of the sequence-to-sequence architecture for regression allows us to utilize all of the pre-trained weights from mT5.

The maximum input sequence length to our metric is 1024 SPM tokens (Kudo and Richardson, 2018). The inputs are truncated during training or inference if the input is larger than 1024.[4] In the worst case, this happens up to 27% of the time on the MQM data for 10 sentences per paragraph (see Appendix A for specific statistics.)

Sentence-Level Baseline. In addition to the paragraph-level metrics, we train a sentence-level version that is trained on the same DA data but only $k = 1$ sentences per paragraph. This baseline metric can be used to directly compare to the paragraph-level metrics that we train because the model architecture, training procedure, etc., are identical. The only difference is the training data. This metric is referred to as SENT-BASE.

Other Metrics. In addition to the metrics described in this paper, we evaluate BLEU (Papineni et al., 2002), COMET-22 (Rei et al., 2020, 2022), and PaLM-2 from Fernandes et al. (2023) as sentence-level metrics applied to paragraphs (i.e., §4.1) and document-level metric BlonDE (Jiang et al., 2022). BLEU scores translations using lexical n-gram overlap, and COMET-22 is a learned regression metric that first embeds the input hypothesis, reference, and source, combines them to a joint representation, then finally predicts a score.

The metric from Fernandes et al. (2023) is based on the PaLM-2 large language model (Anil et al., 2023). We evaluate both the zero shot version, in which PaLM-2 is prompted to score a translation on a scale from 0 to 100, and the regression version that finetunes PaLM-2 on MQM ratings to predict a floating point quality score, similar to COMET. Our analysis includes the Bison variant of PaLM-2.

BlonDE evaluates discourse phenomena in document translations via a set of automatically extracted features. It was designed to evaluate texts longer than paragraphs, like book chapters, but we compare against it in this work. BlonDE is available in English only.

We use the SacreBLEU (Post, 2018) implementation of BLEU and the `Unbabel/wmt22-comet-da` COMET-22 model that was trained on sentence-level WMT DA data from 2017-2020.[5]

[4] We experimentally saw no benefit from removing sequences longer than 1024 tokens during training.

[5] Note that the COMET-22 scores we report come from

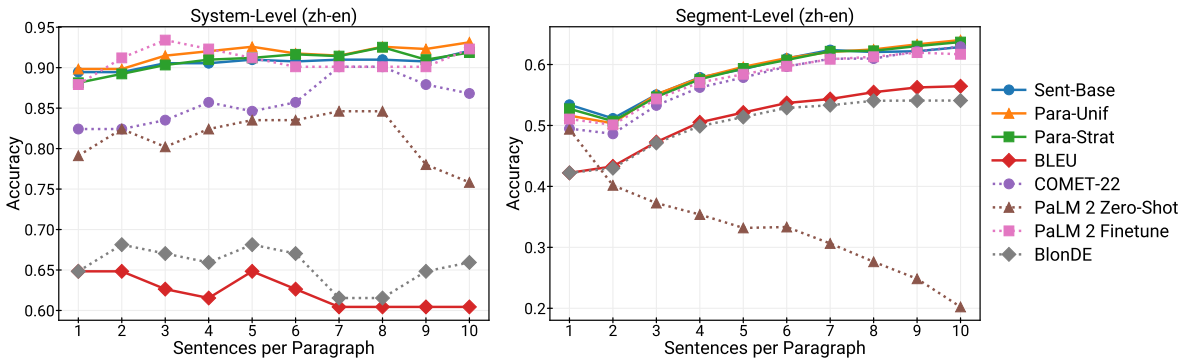


Figure 3: As the number of sentences per paragraph increases, the pairwise accuracy scores (y-axis) of the metrics appears to either not decrease (system-level, left) or increase (segment-level, right). This suggests that accurately scoring a paragraph is an easier task than an individual sentence, even for metrics that are not trained on paragraph-level examples. The results of metrics trained in this work presented here are an average of 5 different runs. Results for other language pairs follow the same trend and are included in Appendix B.

5.3 Meta-Evaluation Metrics

The quality of an evaluation metric is quantified by measuring the correlation of its scores to human ratings of translation quality, a process known as meta-evaluation. In this work, we meta-evaluate metrics using pairwise accuracy at both the system and segment levels.[6] A brief overview of how these accuracy statistics are calculated follows.

At the system-level, an automatic metric and human score is calculated per system by averaging scores over paragraphs. The system-level pairwise accuracy is then computed by enumerating all possible pairs of systems and then calculating the proportion of those pairs for which the automatic metric and human ground-truth ratings agree on their ranking (Kocmi et al., 2021). Thus, the accuracy score can be interpreted as the proportion of pairs of systems that the metric ranked correctly.

At the segment-level, we report segment-level pairwise accuracy using the group-by-item variant of the segment-level correlation in combination with tie calibration (Deutsch et al., 2023). In contrast to system-level accuracy, the group-by-item segment-level correlation calculates the proportion of pairs of *translations* of the same source segment that the metric ranks correctly, then averages that accuracy score over all source segments. The segments used in this evaluation are paragraphs, thus the interpretation of this accuracy score is the proportion of pairs of translations of the same source paragraph that are ranked correctly by the metric.

Because humans frequently assign the same score to translations and regression-based evaluation metrics almost never predict two translations are tied, we follow Deutsch et al. (2023) and run tie calibration before calculating the segment-level accuracy. This procedure automatically introduces ties in the metrics' scores by searching for an ϵ difference in metric score that, when two translations are considered to be a tie if they differ by less than ϵ, achieves the highest accuracy score. We report the accuracy score that corresponds to the best ϵ.

Results using Pearson's correlation follow similar trends to the accuracy results and are available in Appendix B.

5.4 Implementation Details

Our learned metrics are implemented with TensorFlow (Abadi et al., 2015) in the T5X library (Roberts et al., 2022). They are initialized with the XXL version of mT5, which contains 13B parameters. It is trained for a maximum of 20k steps and a batch size of 128 using Adafactor (Shazeer and Stern, 2018) on 64 v3 TPUs. Checkpoint selection was done by selecting the step that has the highest average segment-level pairwise accuracy across language pairs and all values of k sentences per paragraph after applying tie calibration. In general, we observed the specific checkpoint selection strategy was not too important.

only the reference-based regression model, not the ensemble that was submitted to the WMT'22 metrics shared task.

[6]The segment-level correlation could be referred to as a paragraph-level correlation in this work because the segments we evaluate on are paragraphs. However, to be consistent with the evaluation literature, we still use the term segment-level correlation.

6 Results

First, we directly evaluate how well metrics perform when used to directly score paragraphs (§6.1), then we further examine the behavior of different paragraph-level metrics by analyzing their performances with the context of their sentence-level counterparts (§6.2).

6.1 Paragraph-Level Evaluation

Figure 3 plots the system- and segment-level correlation results for different numbers of k sentences per paragraph. Each metric is used to directly score a full paragraph even if the metric was not designed to do so (e.g., SENT-BASE or COMET-22). There are several interesting observations.

Paragraph-Level Performance. First, as the length of the paragraphs increases, the system-level correlations remain relatively steady or increase and the segment-level correlations clearly improve for all metrics, except for PaLM-2 zero-shot. This is evidence that scoring paragraphs is an easier task than scoring individual sentences, a result that is counterintuitive; scoring more text should seemingly be a harder task. We hypothesize this result is explained by the fact that some noise in the human and metric scores is averaged away, leaving more reliable signals as the paragraphs get longer. If the metric scores are unbiased estimators, their agreement with human rating should then increase.

PaLM-2 zero-shot is an outlier in this case because it predicts a large number of ties between translations. Prompting large language models for MT evaluation is known to result in the model predicting a small number of unique scores, resulting in many ties (Kocmi and Federmann, 2023; Fernandes et al., 2023). As the length of the paragraph increases, the number of MQM ties decreases. Since pairwise accuracy penalizes incorrect tie predictions, the zero shot model has worse performance on longer texts. See Figure 4 for a visualization of the number of ties in the PaLM-2 output and MQM scores.

Sentence vs. Paragraph Level. Then, there appears to be little evidence that training on paragraph-level examples results in better correlations to human ratings on paragraph-level test data. For instance, increasing the weight of the paragraph-level data during training does not help compared to uniformly sampling data (compare PARA-STRAT to PARA-UNIF). Further, the baseline metric SENT-BASE that shares the same architecture as our paragraph-level metrics but is only trained on sentence-level data ($k = 1$) performs just as well as the paragraph-level metrics. This observation is additionally supported by COMET-22's results. The difference between the metrics we train versus COMET is relatively constant for all values of k, demonstrating that COMET is not systematically worse on longer inputs.

The generalization of sentence-level metrics on paragraph-level data is rather surprising. The length of the inputs for scoring paragraphs is up to 10x longer than those for scoring sentences (see Table 1). Even though the length of the test data is out-of-distribution with respect to the training data, the sentence-level metrics predict reliable scores on the paragraph-level data. Next, we further analyze the sentence-level metrics to better understand their scores.

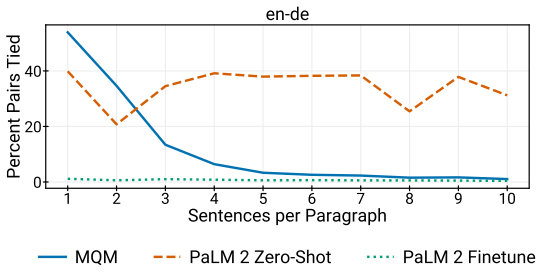


Figure 4: There are fewer MQM ties as the number of sentences per paragraph increases. The finetuned PaLM-2 model outputs a very small number of ties, whereas the zero-shot model consistently predicts a large number of ties. Since the pairwise accuracy meta-evaluation metric penalizes metrics for incorrect tie predictions, the zero-shot model will have worse performance as the inputs get longer.

Dataset	1 Sent. per Para.			10 Sent. per Para.		
	25th	50th	75th	25th	50th	75th
WMT'19 DA	20	31	47	300	362	431
WMT'20 DA	24	38	58	318	410	524
WMT'21 MQM	28	41	57	370	433	516
WMT'22 MQM	15	27	43	265	333	426

Table 1: The SPM token lengths for the given percentiles are in general around 10 times larger with 10 sentences per paragraph compared to a single sentence. Visualizations of the distributions for every paragraph length can be found in Appendix A.

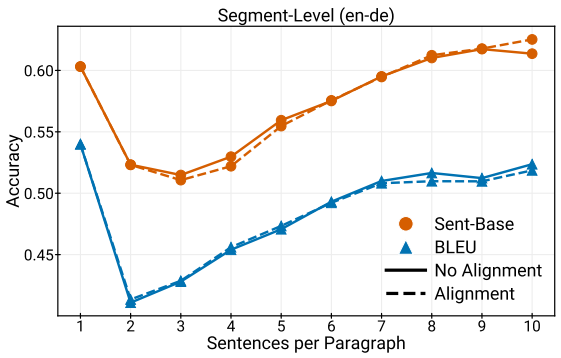


Figure 5: Metrics that score a paragraph directly (solid line) versus those that assume an alignment between the reference and hypothesis and calculate a score by averaging across the k sentence-level values (dashed line) perform very similarly. The drop from 1 sentence to 2 sentences per paragraph is likely due to the fact that a large number of ties in the ground-truth get broken, so introducing more ties via tie calibration is less helpful since doing so is right less often. This phenomenon does not happen with Pearson correlations (see Appendix B).

6.2 Understanding Sentence-Level Metrics

To further analyze the performance of the sentence-level metrics on paragraph-level data, we compare the two versions of applying a sentence-level metric to paragraphs discussed in §4.1. One version directly scores a full paragraph (thus, making no assumption about an alignment between the hypothesis and reference), whereas the other averages the scores of evaluating the individual k hypothesis sentences against the corresponding reference sentence (thus, assuming a sentence-level alignment exists).

Figure 5 shows that for two sentence-level metrics, the baseline trained in this work and BLEU, the performance of the two paragraph scoring variants is very similar. Then, Figure 6 shows that the Pearson correlation between the scores for those two variants is very high (≥ 0.85).

Together, these results point to the fact that there is little difference between these two methods. Directly scoring a paragraph or scoring individual sentences yield both similar scores and similar agreement to human ratings. The sentence-level metrics appear to be scoring full paragraphs in a desirable way—by calculating some average score across sentences.

This result is not obvious. As the length of the input increases, the bag-of-n-grams representation used by lexical matching metrics like BLEU

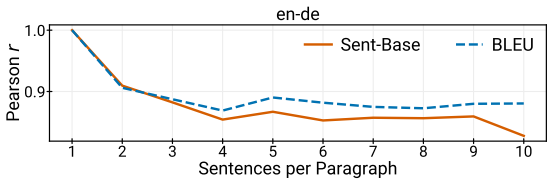


Figure 6: The plot shows the Pearson correlation on en-de between directly predicting a score for a paragraph of k sentences and calculating a paragraph score by averaging over k sentence-level scores. The correlations are quite high, demonstrating that the both methods result in very similar scores.

have an increased potential for erroneous matches between the hypothesis and reference sentences, which could result in misleading scores. Learned metrics, like the ones trained in this work, have not been trained on a significant amount of very long data, so it is not clear that the scoring functions they learn would generalize well to longer inputs. Despite this, the sentence-level metrics appear to predict high-quality scores for paragraphs.

In Section 7, we propose a hypothesis for why this is the case and why training on paragraph-level data does not appear to result in a better metric.

6.3 Literary Translation Evaluation

We compared how frequently SENT-BASE and PARA-STRAT agree with the 540 pairwise human preference judgments between paragraph literary translations from Karpinska and Iyyer (2023). We found that the two models agreed 285 and 305 times, respectively. While it is a positive signal that the paragraph-level model appears to be better aligned with human preferences of longer texts, the difference was not quite statistically significant under a pairwise permutation test with $\alpha = 0.05$ ($p = 0.09$). Future work should perform a more in-depth analysis of this data and collect a larger number of paragraph translations and judgments.

7 Discussion

In theory, training on paragraph-level data should have advantages compared to training on sentence-level data. The metric (1) should be able to handle longer input sequences, (2) it should be able to capture long range dependencies, and (3) it should be able to model different paragraph-level phenomena like information or sentence reordering. However, we were not able to demonstrate these advantages in practice, and we theorize why as follows.

Source Context: Maria said no.
Source: She did not slap the green witch.

Reference Context: Maria dijo no.
Reference: No le dió una bofetada a la bruja verde.

Hypothesis: Ella(✓)/Él(✗) no le dió una bofetada a la bruja verde.

Figure 7: An English-to-Spanish translation example where the reference translation does not have enough information to correctly evaluate the hypothesis. Gender in Spanish is marked on pronouns, and Spanish is a pro-drop language, which means the pronoun can be omitted if the context is clear. In this example, the pronoun is dropped from the reference, so determining whether the pronoun used in the hypothesis requires taking into account the previous reference sentence. We suspect such examples are not frequent, and if they do exist, the information required to resolve the ambiguity is relatively local to the reference sentence.

First, the analysis in §6.2 shows that sentence-level metrics generalize well to significantly longer input, so advantage (1) may not be so relevant. We hypothesize that the scoring function learned by sentence-level metrics like SENT-BASE or COMET could score a token in the hypothesis based on some alignment to the reference using its relative position in the translation. This function would be agnostic with respect to the global positioning, and thus the scoring function would generalize well to longer inputs. If this were true, training on paragraph-level data would not be necessary to obtain good performance on long sequences.

Second, evaluating translation quality seems to be a very "local" problem in the sense that modeling long range dependencies is not frequently necessary for evaluation. Often, the reference phrase that aligns to a hypothesis phrase has enough information to accurately evaluate the hypothesis. If it does not, the information is likely nearby, not several sentences away (see Figure 7). Although the sentence-level metrics were not trained on multiple sentences, we suspect they are able to capture nearby dependencies across sentences when evaluating paragraphs. In theory, a paragraph-level metric would have the ability to model long range dependencies since it could observe them during training. However, if they are infrequent, advantage (2) over sentence-level metrics may be small.

Finally, the ability for our learned paragraph metrics to capture phenomena like sentence reordering is limited by our dataset construction method. Since the paragraphs in our training and test sets come from MT systems that translated one sentence at a time, there are no phenomena like sentence reordering present in the datasets. Therefore, the paragraph-level metric cannot learn to model such cases, and the metrics are never evaluated on them either. Thus, the limitations of the dataset mean that we cannot demonstrate advantage (3).

We believe that paragraph-level metrics are necessary for evaluating true paragraph translations, where MT systems can be more creative with how a full paragraph is translated, rather than paragraph translations that are created by translating individual sentences. We hypothesize that sentence-level metrics will not generalize well when there is no sentence alignment or there is significant information reordering. To accurately evaluate actual paragraph translations, metrics need to be trained on similar data. Future work should invest in collecting human ratings for paragraph-level translations so that new metrics can be trained and evaluated.

8 Related Work

The vast majority of research on MT evaluation has worked at the sentence level (Papineni et al., 2002; Banerjee and Lavie, 2005; Snover et al., 2006; Popović, 2015, 2017; Lo, 2019; Sellam et al., 2020; Rei et al., 2020, 2022; Thompson and Post, 2020; Wan et al., 2022), although there has been recent interest in moving beyond sentence-level evaluation. Vernikos et al. (2022) propose a method to incorporate document-level context into a sentence-level metric by using the additional context when computing the representations for the hypothesis and reference sentences. Although they use document context in their metric, it is still scores single sentences at a time, in contrast to the paragraph-level metrics in our work that predict a score for entire paragraphs at once. Then Jiang et al. (2022) propose a document-level metric called BlonDE that targets evaluating discourse phenomena as opposed to overall translation quality (i.e., they do not model translation accuracy errors). To the best of our knowledge, ours is the first study aimed at training a learned metric that directly scores entire paragraphs.

Other studies that have evaluated sentence-level metrics beyond the sentence-level have done so in the literary domain. Thai et al. (2022) show that automatic metrics prefer MT output over human translations, and Karpinska and Iyyer (2023) show

that metrics prefer actual translations of paragraphs over sentence-by-sentence translations. Our work is complementary to theirs as we focus on the news domain, train metrics on paragraph-level data, and evaluate on a much larger set of human ratings. It is not clear whether conclusions reached about metrics in the news domain will apply to the literary domain or vice versa.

Some researchers have developed challenge sets that can be used to probe how well metrics capture discourse phenomena that appear when translating more than one sentence at a time (Bawden et al., 2018; Müller et al., 2018; Lopes et al., 2020). However, these challenge sets can be trivial for reference-based metrics because the reference often resolves the ambiguity in the translation. To the best of our knowledge, a challenge set that forces reference-based metrics to use context outside of a single reference sentence during evaluation (see Figure 7) does not exist.

Research on generating translations of text longer than single sentences directly use sentence-level metrics to score translations (Tiedemann and Scherrer, 2017; Miculicich et al., 2018; Ma et al., 2020; Wu et al., 2023; Post and Junczys-Dowmunt, 2023). Our work can be viewed as a justification for doing so.

9 Conclusion

In this work, we proposed a method for constructing paragraph-level datasets for training and meta-evaluating MT evaluation metrics from sentence-level data. Our experimental results showed that metrics trained on paragraph-level data do not necessarily out-perform those trained on sentence-level data, potentially due to the fact that sentence-level metrics seem to generalize well to longer inputs and limitations of our paragraph-level datasets. Future work should invest in collecting human judgments for paragraph translations generated by MT systems that directly translate full paragraphs instead of translating one sentence at a time. Such a dataset would be more likely to contain phenomena that do not exist at the sentence level, which we hypothesize would be more likely to require metrics designed to work at the paragraph level.

Limitations

There are a couple of limitations related to our dataset construction approach that are worth enumerating.

As discussed in Section 7, our ability to evaluate metrics' performances on all types of paragraph-level translations is limited by our dataset construction method. Our translated paragraphs are generated by MT systems that translate one sentence at time, which results in sentence aligned data. Therefore, we are unable to evaluate metrics on true paragraph-level translations that might have sentence or information reordering.

Then, the WMT data no longer contains information about the white space between the original source sentences. Therefore, the DA and MQM paragraph-level datasets do not contain the paragraph breaks that were in the original document. Each of the k sentences is concatenated together and separated by a space in our work, so it is very likely that the artificially constructed paragraphs do not perfectly resemble real paragraphs.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang

Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. PaLM 2 Technical Report.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 Conference on Machine Translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating Discourse Phenomena in Neural Machine Translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties Matter: Meta-Evaluating Modern Metrics with Pairwise Accuracy and Tie Calibration.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André F. T. Martins, Graham Neubig, Ankush Garg, Jonathan H. Clark, Markus Freitag, and Orhan Firat. 2023. The Devil is in the Errors: Leveraging Large Language Models for Fine-grained Machine Translation Evaluation.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 Metrics Shared Task: Stop Using BLEU – Neural Metrics Are Better and More Robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 Metrics Shared Task: Evaluating Metrics with Expert-based Human Evaluations on TED and News Domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An Automatic Evaluation Metric for Document-level Machine Translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States. Association for Computational Linguistics.

Yuchen Eleanor Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Mrinmaya Sachan, and Ryan Cotterell. 2023. Discourse-Centric Evaluation of Document-level Machine Translation with a New Densely Annotated Parallel Corpus of Novels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7853–7872, Toronto, Canada. Association for Computational Linguistics.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist.

Tom Kocmi and Christian Federmann. 2023. Large Language Models Are State-of-the-Art Evaluators of Translation Quality.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing.

In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Chi-kiu Lo. 2019. YiSi - a Unified Semantic MT Quality Evaluation and Estimation Metric for Languages with Different Levels of Available Resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica*, (12):0455–463.

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level Neural MT: A Systematic Comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 Metrics Shared Task: Segment-Level and Strong MT Systems Pose Big Challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Shuming Ma, Dongdong Zhang, and Ming Zhou. 2020. A Simple and Effective Unified Encoder for Document-Level Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3505–3511, Online. Association for Computational Linguistics.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 Metrics Shared Task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-Level Neural Machine Translation with Hierarchical Attention Networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A Large-Scale Test Set for the Evaluation of Context-Aware Pronoun Translation in Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Matt Post and Marcin Junczys-Dowmunt. 2023. Escaping the sentence-level paradigm in machine translation.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Adam Roberts, Hyung Won Chung, Anselm Levskaya, Gaurav Mishra, James Bradbury, Daniel Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, Curtis Hawthorne, Aitor Lewkowycz, Alex Salcianu, Marc van Zee, Jacob Austin, Sebastian Goodman, Livio Baldini Soares, Haitang Hu, Sasha Tsvyashchenko, Aakanksha Chowdhery, Jasmijn Bastings, Jannis Bulian, Xavier Garcia, Jianmo Ni, Andrew Chen, Kathleen Kenealy, Jonathan H. Clark, Stephan Lee, Dan Garrette, James Lee-Thorp, Colin Raffel, Noam Shazeer, Marvin Ritter, Maarten Bosma, Alexandre Passos, Jeremy Maitin-Shepard, Noah Fiedel, Mark Omernick, Brennan Saeta, Ryan Sepassi, Alexander Spiridonov, Joshua Newlan, and Andrea Gesmundo. 2022. Scaling Up Models and Data with `t5x` and `seqio`. *arXiv preprint arXiv:2203.17189*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. Rethinking Document-level Neural Machine Translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3537–3548, Dublin, Ireland. Association for Computational Linguistics.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring Document-Level Literary Machine Translation with Parallel Paragraphs from World Literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020. Automatic Machine Translation Evaluation in Many Languages via Zero-Shot Paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Jörg Tiedemann and Yves Scherrer. 2017. Neural Machine Translation with Extended Context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Zhaopeng Tu, Yang Liu, Shuming Shi, and Tong Zhang. 2018. Learning to Remember Translation History with a Continuous Cache. *Transactions of the Association for Computational Linguistics*, 6:407–420.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly Easy Document-Level MT Metrics: How to Convert Any Pretrained Metric into a Document-Level Metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 2022. UniTE: Unified Translation Evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127, Dublin, Ireland. Association for Computational Linguistics.

Minghao Wu, George Foster, Lizhen Qu, and Gholamreza Haffari. 2023. Document Flattening: Beyond Concatenating Context for Document-Level Neural Machine Translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 448–462, Dubrovnik, Croatia. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

A Dataset Statistics

The exact number of paragraph-level instances by WMT year and language pair that we generaetd from our dataset construction procedure (see §3) can be found in Table 2 for DA and Table 3 for MQM. Figure 8 visualizes the distribution of the lengths of the hypotheses in the paragraph-level datasets based on mT5 SPM tokens. Then, Table 4 contains the number of paragraph examples that are too long to fit into the 1024 SPM maximum context length that is used by the metrics trained in this work.

B Additional Results

Figure 9 contains the system- and segment-level accuracy correlations on the en-de and en-ru language pairs from WMT'22 MQM that were not presented in the main body of the paper. Figure 10 contains the correlations for all 3 language pairs but uses Pearson correlation instead of pairwise accuracy.

Figure 11 shows the correlation between the two ways to apply a segment-level metric to paragraph-level data, directly scoring the paragraph or averaging the k segment scores, on the en-ru and zh-en WMT'22 MQM dataset.

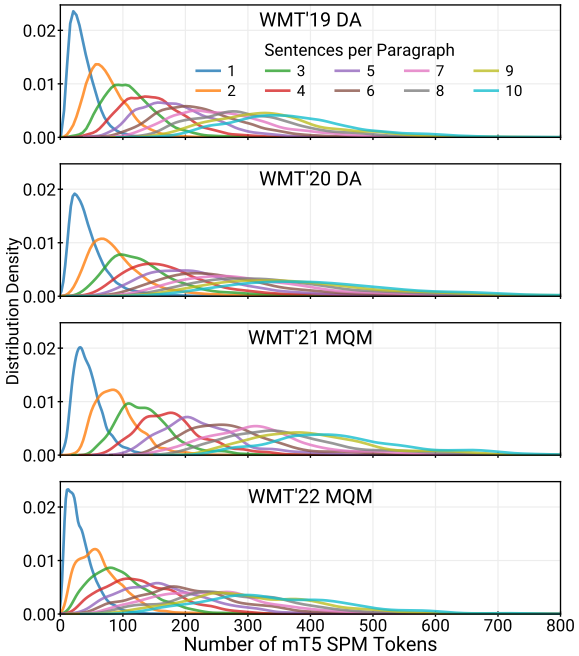


Figure 8: The distribution of the length of the hypothesis translations for the direct assessment (DA) and MQM datasets for a given number of sentences per paragraph.

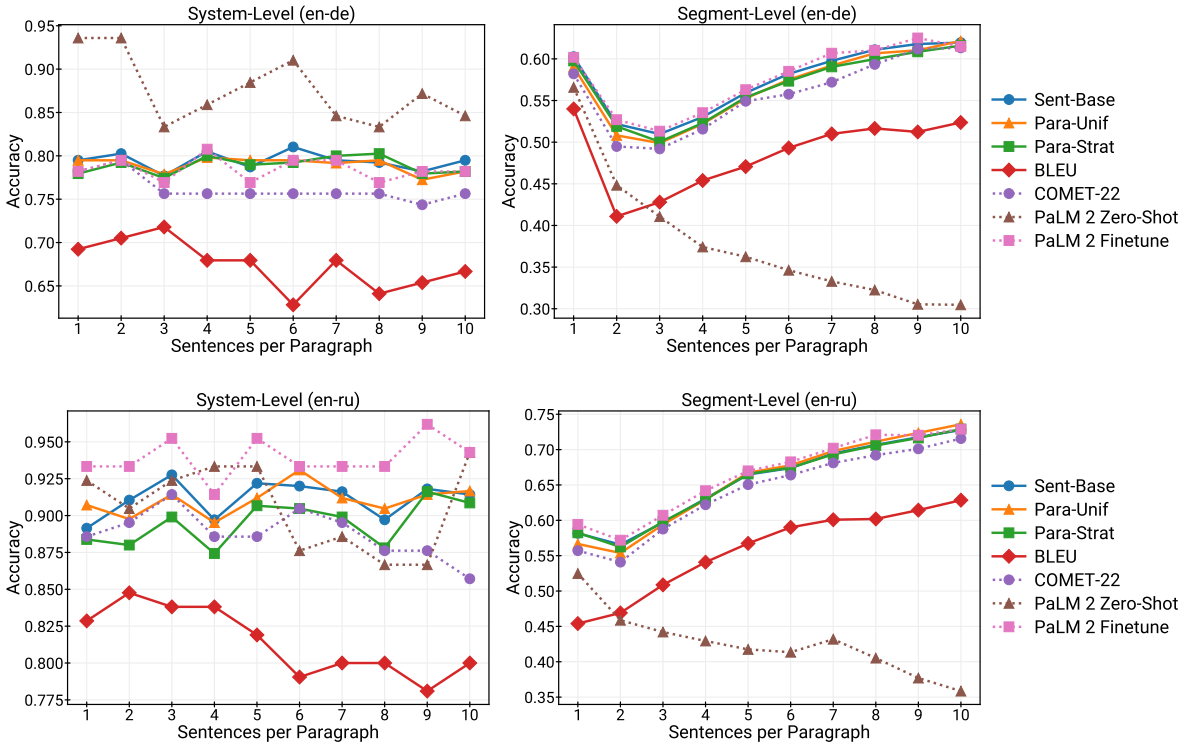


Figure 9: System- and segment-level accuracy results for the en-de and en-ru language pairs on the paragraph-level WMT'22 MQM data for different numbers of k sentences per paragraph. In general, the system-level correlations are relatively flat and the segment-level correlations increase as the number of sentences per paragraph increases. BlonDE is not included because it only supports English.

Year	LP	Sentences per Paragraph									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
2019	de-cs	16900	1032	95	12	1	0	0	0	0	0
2019	de-en	34756	16754	10896	7735	5976	4660	3947	3147	2730	2345
2019	de-fr	6700	173	5	0	0	0	0	0	0	0
2019	en-cs	27445	13241	8710	6152	4834	3865	3215	2607	2371	1967
2019	en-de	45131	21777	14311	10124	7932	6363	5274	4285	3906	3232
2019	en-fi	20618	9937	6557	4611	3628	2910	2419	1945	1799	1482
2019	en-gu	10151	4890	3229	2267	1774	1423	1221	964	884	722
2019	en-kk	12922	6221	4115	2888	2253	1813	1562	1223	1112	910
2019	en-lt	13217	6363	4219	2963	2319	1863	1603	1257	1137	944
2019	en-ru	22600	10902	7180	5069	3974	3185	2650	2137	1966	1633
2019	en-zh	26530	12810	8434	5944	4673	3758	3102	2520	2308	1904
2019	fi-en	20286	362	21	2	0	0	0	0	0	0
2019	fr-de	4000	87	3	0	0	0	0	0	0	0
2019	gu-en	14860	550	40	2	0	0	0	0	0	0
2019	kk-en	15763	705	77	10	0	0	0	0	0	0
2019	lt-en	16046	489	32	2	0	0	0	0	0	0
2019	ru-en	24247	785	83	10	1	0	0	0	0	0
2019	zh-en	50722	15164	9347	6774	5030	4087	3312	2714	2226	1797
2020	cs-en	9381	4322	2628	1797	1323	940	685	404	241	138
2020	de-en	12541	5825	3451	2422	1808	1220	927	652	507	378
2020	en-cs	34180	16371	10324	7358	5591	4501	3474	2749	2270	2035
2020	en-de	17393	8337	5253	3723	2859	2283	1729	1362	1138	1033
2020	en-iu	6145	3028	1990	1479	1152	937	801	693	600	538
2020	en-ja	21999	10672	6769	5036	3907	3093	2513	2109	1812	1635
2020	en-pl	18342	8891	5636	4192	3266	2569	2089	1756	1514	1377
2020	en-ru	19543	9494	6058	4433	3477	2750	2279	1847	1602	1468
2020	en-ta	9175	4439	2825	2100	1634	1301	1035	875	746	680
2020	en-zh	41965	20069	12656	9034	6843	5510	4260	3371	2782	2483
2020	iu-en	12172	75	0	0	0	0	0	0	0	0
2020	ja-en	9879	4710	3047	2103	1715	1321	1053	845	759	639
2020	km-en	6951	72	0	0	0	0	0	0	0	0
2020	pl-en	12435	6048	3871	2857	2184	1708	1445	1265	1030	844
2020	ps-en	7138	110	2	0	0	0	0	0	0	0
2020	ru-en	11244	5369	3408	2405	1832	1488	1179	952	785	604
2020	ta-en	7842	3762	2406	1723	1322	1065	847	694	572	473
2020	zh-en	30325	14567	9253	6674	5106	4078	3374	2811	2223	1824

Table 2: The number of paragraphs with the given number of sentences per paragraph from the direct assessment data from WMT'19 and WMT'20. Each paragraph is required to a contiguous block of sentences that are rated by the same rater.

Dataset	LP	Sentences per Paragraph									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
WMT'21	en-de	7905	3825	2460	1800	1395	1140	870	765	660	585
WMT'21	en-ru	7905	3825	2460	1800	1395	1140	870	765	660	585
WMT'21	zh-en	9058	4340	2814	1974	1596	1190	994	770	658	644
WMT'22	en-de	18410	8932	5236	3486	3080	1610	1568	1470	1372	1330
WMT'22	en-ru	19725	9570	5610	3735	3300	1725	1680	1575	1470	1425
WMT'22	zh-en	28110	13005	7935	5655	4245	3285	2670	2160	1935	1710

Table 3: The number of paragraphs with the given number of sentences per paragraph from the MQM data from WMT'21 and WMT'22. Each paragraph is required to a contiguous block of sentences that are rated by the same rater.

Dataset	Sentences per Paragraph									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
WMT'19 DA	2 (0%)	3 (0%)	4 (0%)	15 (0%)	48 (0%)	196 (1%)	440 (2%)	702 (3%)	1349 (7%)	1944 (11%)
WMT'20 DA	4 (0%)	179 (0%)	667 (1%)	1148 (2%)	1598 (4%)	2222 (6%)	2879 (10%)	3389 (15%)	4041 (22%)	4688 (29%)
WMT'21 MQM	0 (0%)	0 (0%)	0 (0%)	0 (0%)	3 (0%)	23 (1%)	103 (4%)	245 (11%)	295 (15%)	488 (27%)
WMT'22 MQM	0 (0%)	0 (0%)	6 (0%)	11 (0%)	56 (1%)	74 (1%)	110 (2%)	202 (4%)	266 (6%)	450 (10%)

Table 4: The number (and percent) of paragraphs for which the number of SPM tokens in the reference and hypothesis combined is larger than the maximum allowable input length by our metric, 1024. If the input is too long, it is truncated.

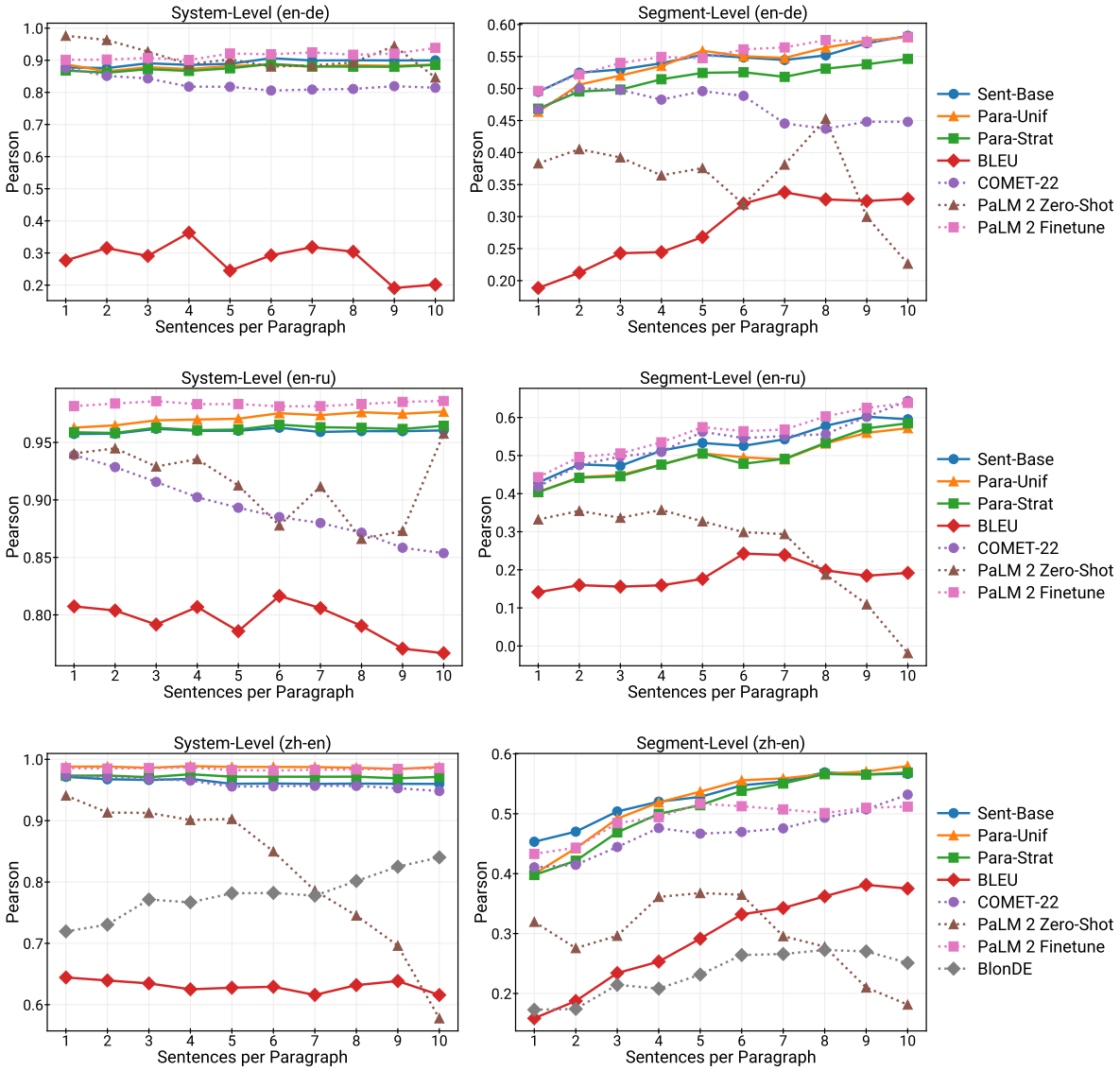


Figure 10: The system- and segment-level correlation results when using Pearson correlation follow very similar trends to those that use pairwise accuracy. The segment-level Pearson uses the "no grouping" variant from Deutsch et al. (2023) to avoid the NaN problem that happens with the "group-by-item" variant, which was used in combination with pairwise accuracy in the main body of the paper.

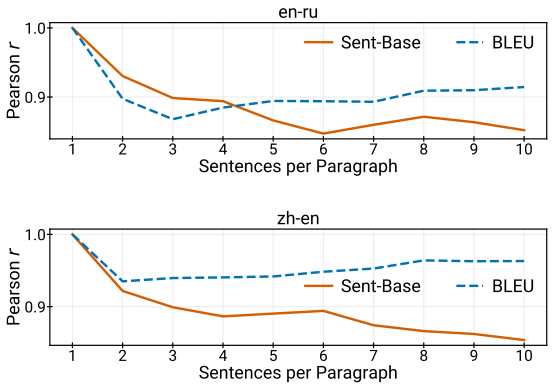


Figure 11: The correlation between metric scores for directly scoring paragraphs and averaging the score of evaluating the k sentences per paragraph independently on the WMT'22 MQM data.

Automating Behavioral Testing in Machine Translation

Javier Ferrando[◇*] **Matthias Sperber**[†] **Hendra Setiawan**[†]
Dominic Telaar[†] **Saša Hasan**[†]
[◇]Universitat Politècnica de Catalunya
[†]Apple
javier.ferrando.monsonis@upc.edu,sperber@apple.com

Abstract

Behavioral testing in NLP allows fine-grained evaluation of systems by examining their linguistic capabilities through the analysis of input-output behavior. Unfortunately, existing work on behavioral testing in Machine Translation (MT) is currently restricted to largely handcrafted tests covering a limited range of capabilities and languages. To address this limitation, we propose to use Large Language Models (LLMs) to generate a diverse set of source sentences tailored to test the behavior of MT models in a range of situations. We can then verify whether the MT model exhibits the expected behavior through matching candidate sets that are also generated using LLMs. Our approach aims to make behavioral testing of MT systems practical while requiring only minimal human effort. In our experiments, we apply our proposed evaluation framework to assess multiple available MT systems, revealing that while in general pass-rates follow the trends observable from traditional accuracy-based metrics, our method was able to uncover several important differences and potential bugs that go unnoticed when relying only on accuracy.[1]

1 Introduction

Automatic evaluation metrics such as BLEU (Papineni et al., 2002) and COMET (Rei et al., 2020) are the primary means of measuring the translation quality of MT systems. Researchers and practitioners rely on them for comparing systems, detecting regressions, and making deployment decisions. This poses an important concern: such metrics typically aggregate the performance of systems across a set of sentences into single scores. Unfortunately, these metrics by design tend to overlook specific infrequent but important error cases, making it difficult to reliably detect such issues in practice.

Property	Translation Errors
Integers	7000000 → 70.000.000
Decimals	500.75 → 500.75
Large Numbers	1.366 billion → 1.366 Milliarden
Idioms	ins and outs → Ins und Outs
Currencies	BRL → RL
Physical Units	miles → km
Web Terms	www.onlinegrocery.com → www.onlineegrocery.com
...	...

Table 1: Subset of linguistic properties tested with our proposed method, and examples (source → translation) of translation errors found in En→De MT models.

Behavioral testing, originally developed as a type of software testing (Beizer and Wiley, 1996), has been proposed as an approach that can alleviate such kinds of problems in natural language processing (Ribeiro et al., 2020). Behavioral tests focus on assessing a system's fine-grained linguistic capabilities by validating input-output behavior in a controlled fashion.

Table 1 shows examples of typical issues of MT systems that could be covered by behavioral tests. We argue that the availability of a comprehensive behavioral test suite for MT would be of high practical value: It would allow understanding how exactly two MT models differ, or to block an MT system from being deployed if a passing threshold for a certain linguistic capability is not met.

However, there are currently two major limitations that arise when attempting to apply behavioral testing to MT. First of all, behavioral testing was originally designed for evaluating systems characterized by a relatively small output space. For instance, Ribeiro et al. (2020) investigate sentiment classification, duplicate question detection, and machine comprehension. In contrast, the output space of MT systems grows exponentially as tokens are generated. Secondly, behavioral testing often requires rigid templates to create examples and their

[*] Work done during an internship at Apple.
[1]Prompts and generated data are available at https://github.com/apple/ml-behavioral-testing-for-mt.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 1014–1030
December 6–7, 2023.

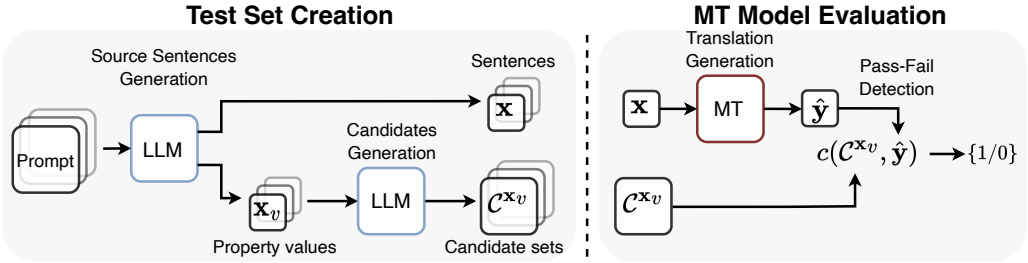


Figure 1: Pipeline of the proposed approach. Left: For each property type, a test set is created via a LLM, composed of source sentences $\mathbf{x}$ with property values $\mathbf{x}_v$ (§3). Subsequently, a candidate set of valid translations of each property value $\mathcal{C}^{\mathbf{x}_v}$ is generated (§4). Right: During evaluation, the translation $\hat{\mathbf{y}}$ generated by an MT model is compared against the candidate sets, and a pass-fail decision is made (§5).

corresponding labels, which involves a costly human effort to develop and expand to additional use cases. Otherwise, the diversity of sentences in the resulting test suite is too limited.

Several recent works have partially addressed these limitations. For example, Wang et al. (2021); Raunak et al. (2022) propose MT-specific test sets which include the ability to handle large output space. Yang et al. (2022) address the limitation of rigid templates. To the best of our knowledge, no prior work has addressed both limitations for MT.

In this study, we aim to bridge this gap by leveraging LLMs with in-context learning to automate the creation of behavioral tests in MT for the first time. Our main contributions are as follows:

- We use LLMs to automate the generation of a diverse set of source sentences for behavioral testing. Sentences are generated to exhibit the specific language property that is being tested.
- We verify whether an MT system's output contains an accurate translation of the language property that is being tested. To this end, we propose using LLMs to generate candidate sets of ground-truth translations of the property values in cases where exhaustive candidate sets are plausible. Otherwise, we generate contrastive candidates and evaluate via semantic similarity measures.
- We present an evaluation framework to robustly compute pass rates of MT models across various language properties, and show results for widely used open-source models on three language pairs.

2 Behavioral Testing for MT

Behavioral testing, as proposed by Ribeiro et al. (2020), uses input-output pairs tailored to evaluate a model capability to correctly handle certain language properties. The goal is to complement traditional aggregated accuracy scores, which, while useful by themselves, often fail to capture long-tail phenomena. In practice, manual inspection of system outputs is often crucial to make up for this shortcoming. Automated behavioral testing provides a more reliable and less cumbersome alternative that can reduce or eliminate the need for manual inspection, provided that a sufficient range of language properties is tested. Test results are presented in the form of a table of pass rates (one pass rate for each tested property) that is informative to decide on consequent steps, e.g. whether bugs must be addressed before deployment. Note that the creation of a *sufficiently comprehensive* behavioral test suite depends crucially on whether its creation can be automated to a high degree, which is also our main design goal in this work.

We are particularly interested in a specific type of behavioral tests, *minimal functionality tests* (MFTs) (Ribeiro et al., 2020).[2] In the context of MT, an MFT measures a model's ability to translate particular property values that appear naturally embedded in some given source sentences.

Figure 1 illustrates our proposed framework. First, a source sentence $\mathbf{x} = \{\boldsymbol{x}_1 \cdots, \boldsymbol{x}_{|\mathbf{x}|}\}$ that contains a tagged property value $\mathbf{x}_v \subseteq \mathbf{x}$ is generated (§3). For instance, if our test property is physical unit translation, we might have $\mathbf{x}$=*"I ran 3 miles."* and $\mathbf{x}_v$=*miles*. A main challenge comes from the fact that there is a potentially large space of correct translations. However, note that by design MFTs only need to check whether the property under test is translated correctly, while un-

[2] Ribeiro et al. (2020) also propose *directional* and *invariance* tests which check how model outputs change under certain input perturbance, but these appear less applicable to MT given the potentially large space of correct translations.

```
You are an assistant that generates sentences where only appears one B = {property}.
Don't be repetitive, change the topic and B between sentences. Write every B inside [].
B must happen only once in each sentence and can only contain {property}.

Write 3 examples.

- {Source sentence demonstration #1}
- {Source sentence demonstration #2}
- {Source sentence demonstration #3}

Now write 10 more diverse sentences itemizing them with '-':
```

Figure 2: General template of the prompt used for generating batches of source sentences.

related translation errors should be ignored. In many cases, this reduces the space of correct translations to a manageable size. We therefore propose to automatically generate a candidate set $\mathcal{C}^{\mathbf{x}_v}$ (either exhaustive or contrastive; see §4) and then apply a pass-fail detector that uses either string matching or semantic similarity measures (§5). In our example, we might generate an exhaustive candidate set $\mathcal{C}^{\mathbf{x}_v}$=$\{$*Meilen*, *mi*$\}$ for the case of translating into German. We now aim to evaluate an MT model $f : \mathbf{x} \mapsto \hat{\mathbf{y}}$. To do so, we compare $\hat{\mathbf{y}}$ against $\mathcal{C}^{\mathbf{x}_v}$. A correct translation $\hat{\mathbf{y}}$=*"Ich lief 3 Meilen."* would match the candidate set and therefore pass the test, while a typical incorrect translation $\hat{\mathbf{y}}$=*"Ich lief 3 km."* does not match the candidate set and therefore fails the test.

Given this general overview of our method, we now turn to a more precise description of each proposed step in the following sections.

3 Source Sentence Generation

To create source sentences for testing a certain language property, we pose several desiderata: Sentences should be *diverse* (e.g. not rely only on a handful of templates), *natural*, *numerous* enough to yield statistical significance, and *contain a property value* associated with our tested property.

Note that existing approaches often struggle to generate diverse test sets due to the reliance on hand-crafted templates (Wang et al., 2021). To overcome this shortcoming, we design a general template for prompting LLMs, in our case ChatGPT[3], OpenAI's model built on InstructGPT (Ouyang et al., 2022). This allows us to generate diverse source language sentences that contain property values suitable for testing different capabilities (see prompt[4] in Figure 2). We instantiate the prompt once for every language property that we wish to include in our test suite.

Candidates Examples
kilometers → kilómetros, km
watts → vatios, W
meters per second → metros por segundo, m/s

Table 2: Examples of En→Es set of candidates generated by ChatGPT.

To simplify the later verification step, we generate sentences that contain exactly one such property value $\mathbf{x}_v$.[5] We generate source sentences with brackets around the property value for easy parsing. A possible test sentence for the property of translating decimal numbers might look as follows:

$$\text{The company received [} \overset{\text{property value}}{\overbrace{\text{4200.4}}} \text{]€.} \qquad (1)$$

Note that brackets are removed before passing the sentence to the MT model.

We apply basic filters to remove duplicated sentences, examples with more than one property value, or those composed of more than one sentence. We repeatedly feed the same prompt to the LLM, and stop the generation process when reaching 1,000 sentences after filtering. Our experiments (§9) indicate that ChatGPT is able to generate sentences of adequate quality and diversity.

4 Candidates Generation

Next, in order to be able to verify whether an MT system correctly translated the property value in the source sentence, we automatically generate valid translation candidate sets for each property value. For some properties, such as number translation, we create exhaustive or near-exhaustive candidate sets. For other properties where the number of valid translations would be too big to do so, we instead

[3] `gpt-3.5-turbo` API accessed on May 2023.

[4] We set temperature=0.9, presence_penalty=2.

[5] For some types of properties, multiple property values may be more appropriate. This is left for future work.

create contrastive candidate pairs that demonstrate desired and undesired behavior. Note that candidate sets only need to be created once and can then be re-used for every tested system.

4.1 Near-Exhaustive Candidate Sets

In this approach, we follow Raunak et al. (2022) in creating a set of all valid translations of each property value in the test (see example in Table 2). However, instead of manually designing candidate sets, we propose using the in-context learning (Brown et al., 2020) and multilingual capabilities of instruction-tuned LLMs (Wei et al., 2022) to accomplish the task. For each property value $\mathbf{x}_v$, we generate a set of translation candidates $\mathcal{C}^{\mathbf{x}_v}$ with ChatGPT (`gpt-3.5-turbo`) (see prompt[6] in Figure 3). We tried to design demonstrations to encompass both correctness and completeness, including possible inflections. An example of demonstrations used for the currencies test can be seen in Appendix B. Note that while we aim for completeness, i.e. all valid translations should be included in the candidate set, in practice we found that some rare translation choices may not be included in the automatically generated candidate sets. However, this will not impact pass-rates much because by nature rare translation choices appear in the MT system's output only in rare situations. In §9 we perform a human assessment of the reliability of the generated candidate sets.

4.2 Contrastive Candidate Pairs

Some property values can span multiple words on the source side, potentially increasing the number of valid translations drastically. An example is idiomatic expressions, where there is an increased risk that the candidate set cannot exhaust all possibilities. To mitigate this issue, we propose using *contrastive* candidate sets an alternative approach.

Given a source property value, we generate a *contrastive* candidate set $\mathcal{C}^{\mathbf{x}_v}_{\text{contra}}$ formed by a correct translation $c^{\mathbf{x}_v}_{\text{corr}}$, and a foil (incorrect) translation $c^{\mathbf{x}_v}_{\text{foil}}$. Appendix C shows an example prompt. Intuitively, an MT model should pass the test sentence if its translation is closer to $c^{\mathbf{x}_v}_{\text{corr}}$ than it is to $c^{\mathbf{x}_v}_{\text{foil}}$.

5 Pass-Fail Detector

Equipped with these candidate sets, we now wish to mark every MT-translated sentence as either *pass* or *fail*. Depending on whether near-exhaustive or contrastive candidate pairs are used, we design pass-fail detectors based on string matching or semantic similarity, respectively.

As it is our goal to design tests that target specific language properties, our pass-fail detectors should only detect cases where the property value under the test is translated incorrectly. Unrelated translation errors should not cause a sentence to be marked as incorrect.[7]

Algorithm 1: Similarity score between translation and contrastive candidate.

Input: $\hat{\mathbf{y}}$: model translation; c: candidate translation; e: encoder
Output: max_sim$(\hat{\mathbf{y}}, c)$
max_sim $\leftarrow -\infty$
$n \leftarrow |c|$
$\mathcal{G}_{\hat{\mathbf{y}}} \leftarrow n$-gram$(\hat{\mathbf{y}}, n)$
$\boldsymbol{c}_{\text{emb}} \leftarrow e(c)$
for $\mathbf{g} \in \mathcal{G}_{\hat{\mathbf{y}}}$ **do**
  $\boldsymbol{g}_{\text{emb}} \leftarrow e(\mathbf{g})$
  if $sim(\boldsymbol{g}_{emb}, \boldsymbol{c}_{emb}) > max\_sim$ **then**
    max_sim $\leftarrow$ sim$(\boldsymbol{g}_{\text{emb}}, \boldsymbol{c}_{\text{emb}})$
return max_sim

5.1 String Matching for Near-Exhaustive Candidate Sets

For the near-exhaustive candidate sets, we define a pass-fail function $c(\hat{\mathbf{y}}, \mathcal{C}^{\mathbf{x}_v}) \in \{0, 1\}$ that takes the model's translation $\hat{\mathbf{y}}$, and the candidates set $\mathcal{C}^{\mathbf{x}_v}$, and returns 1 (pass) if $\hat{\mathbf{y}}$ has a valid translation of the property value, i.e. if it has an element in $\mathcal{C}^{\mathbf{x}_v}$, and 0 (fail) otherwise:

$$c(\hat{\mathbf{y}}, \mathcal{C}^{\mathbf{x}_v}) = \begin{cases} 1 & \text{if } \hat{\mathbf{y}} \cap \mathcal{C}^{\mathbf{x}_v} \neq \emptyset \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Specifically, we consider as pass an exact case-insensitive substring matching. Following Example 1, where $\mathbf{x}_v = 4200.4$, if we are evaluating the En→De decimal numbers translation capabilities of the model, we would consider the model passes the test if it outputs '4200,4', or '4.200,4'.[8]

[6] We use the same set of parameters as for the source sentence generation.

[7] For our purposes, we do not consider whether the translated property is placed at the correct position in the target sentence, but only whether it is correct when considered in isolation. We argue that errors related to fluency and reordering are better evaluated through established accuracy-based metrics.

[8] Note that this involves a design decision: The test case writers must make a decision whether or not the added decimal point is acceptable for their particular use cases.

```
You are a {source_lang}-{target_lang} translator. Given a {property}, write as many valid {target_lang}
translations as you can. Use "|" to separate between valid translations.
Write "NA" if unable to accomplish the task.

{Source property demonstration #1}
{Candidates set source property demonstration #1}
{Source property #2}
{Candidates set source property demonstration #2}
{Source property #3}
{Candidates set source property demonstration #3}

{Source property}
```

Figure 3: General template of the prompt used for generating near-exhaustive sets of candidate translations.

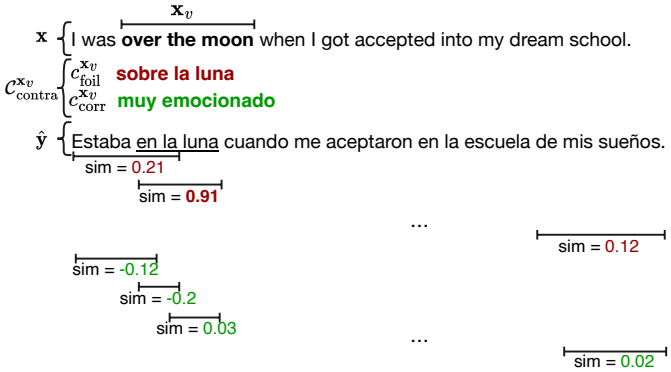


Figure 4: Example of the Contrastive Candidate Pairs approach, where `sim` indicates the semantic similarity between the correct candidate ('muy emocionado') and the 2-grams (in green), and the foil candidate ('sobre la luna') and the 3-grams of the MT translation (in red).

5.2 Semantic Similarity for Contrastive Candidate Pairs

For measuring the closeness of the property value translation to the contrastive candidates, we propose relying on the semantic similarity of word sequences representations extracted by a multilingual encoder (Reimers and Gurevych, 2019, 2020).[9] However, directly measuring the similarity between the translation of the property value and the candidate sets may be unreliable since they may differ in length and the location of the translation is unknown due to lack of word-level alignment. Instead, we propose that, for each candidate $c^{\mathbf{x}_v}_{\text{corr}}$ or $c^{\mathbf{x}_v}_{\text{foil}}$, we split the model's translation into n-grams, where n is the number of words of the current candidate. Then, we measure the similarity between each of the n-grams and the candidates.

Given a translation and the *contrastive* candidate set $\mathcal{C}^{\mathbf{x}_v}_{\text{contra}}$ formed by the correct and foil candidates, we define the pass-fail function as:

$$c(\hat{\mathbf{y}}, \mathcal{C}^{\mathbf{x}_v}_{\text{contra}}) = \begin{cases} 1 & \text{if } \max\_\text{sim}(\hat{\mathbf{y}}, c^{\mathbf{x}_v}_{\text{corr}}) \geq \max\_\text{sim}(\hat{\mathbf{y}}, c^{\mathbf{x}_v}_{\text{foil}}) \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Algorithm 1 formalizes the computation of max_sim function, Figure 4 shows an example.

6 Evaluation Metrics

Having established pass-fail detection for individual sentences, the final step is to compute aggregated *pass rates* across test sets. Appealingly, pass rates are naturally expressed as percentages, making them intuitive to interpret.

6.1 Macro Pass Rate

Let us assume that we have computed pass-fail results across a behavioral test set consisting of N test cases (sentences). From a statistical viewpoint, we have access to a sample $\mathcal{X} = \{c(\hat{\mathbf{y}}^n, \mathcal{C}^{\mathbf{x}^n_v})\}_{n=1}^{N}$, drawn from some unknown distribution over test cases, F. The expectation of the true pass rate can be computed as follows:

$$\text{PR}^{(\mathcal{X})} = \frac{1}{N} \sum_{n}^{N} c(\hat{\mathbf{y}}^n, \mathcal{C}^{\mathbf{x}^n_v}) \tag{4}$$

[9] Employing LLMs is also possible but not explored here because it needs to be applied for every evaluated MT system, incurring higher computational costs.

Model	En→De			En→Es			En→Ja		
	spBLEU	ChrF	COMET-22	spBLEU	ChrF	COMET-22	spBLEU	ChrF	COMET-22
M2M 418M	31.08	57.22	79.49	25.33	51.26	80.63	23.57	32.22	84.84
M2M 1.2B	39.37	62.51	85.35	29.06	53.85	84.22	27.46	35.25	87.63
NLLB 600M	38.88	61.85	85.89	30.65	54.76	85.34	18.75	29.62	86.72
NLLB 3.3B	44.41	65.26	87.98	32.69	56.09	86.39	20.76	32.5	88.12
OPUS MT (Bil)	40.96	63.49	84.61	30.57	54.97	84.9	-	-	-
WMT21 (En-X)	**49.38**	**68.94**	88.76	-	-	-	39.89	44.95	91.95
Commercial system	49.34	68.84	**89.34**	**34.43**	**57.58**	**86.92**	**41.05**	**47.06**	**92.19**

Table 3: Translation scores of the different models used in FLORES-200 devtest set.

One issue that arises in practice is that property values themselves follow a long tail pattern: Certain values appear relatively frequently, while many other values appear only once across the generated test set. This can make pass rates overly sensitive to whether models happen to perform well for these particular values. To mitigate this issue, we assume a generative story in which property values are drawn from a uniform distribution, and consequently compute the expected pass rate as the macro average across property values:

$$\text{MPR}^{(\mathcal{X})} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \frac{1}{N_v} \sum_{i}^{N_v} c(\hat{\mathbf{y}}^i, \mathcal{C}^{\mathbf{x}_v^i}) \quad (5)$$

where $\mathcal{V}$ refers to the set of distinct property values, and N_v to the number of examples associated with each specific property value.

6.2 Confidence Intervals

Although previous work performing behavioral testing for MT shows point estimate scores, confidence intervals provide a more reliable approach to statistical analysis, as they quantify the uncertainty associated with that estimate, and ensure the sample size is large enough. To compute confidence intervals for our estimator MPR we use the Bootstrap method (Efron, 1979), which performs sampling with replacement from $\mathcal{X}$, generating K resamples $\{\mathcal{Y}^1, \cdots, \mathcal{Y}^K\}$, from which we compute their corresponding macro pass rates $\{\text{MPR}^{(\mathcal{Y}^1)}, \cdots, \text{MPR}^{(\mathcal{Y}^K)}\}$ to construct the bootstrap distribution MPR_{boot}. Assuming the distribution of $\mathcal{X}$ is a reasonable approximation of the population distribution F, confidence intervals can be derived from MPR_{boot}. For that purpose, we compute the percentile bootstrap interval for $\alpha = 0.05$ provided by SCIPY library (Virtanen et al., 2020).

6.3 Paired Bootstrap

The paired bootstrap is a statistical resampling technique used to assess the uncertainty and make inferences about the difference between two samples. Paired bootstrap allows us to compare the property's sample of passes/fails for two different models (Koehn, 2004). By following the resampling process outlined in the previous section, if a model consistently outperforms the other in 95% of the iterations, we can assert with 95% statistical significance that it is superior.

7 Properties to Test

We design a number of tests and use our proposed framework to evaluate MT models in multiple properties. The chosen properties, also studied in the literature (Wang et al., 2021; Raunak et al., 2022), have two important qualities that make them useful for evaluating translation systems: vital for producing high-quality translations, yet posing a challenge when assessing through conventional evaluation metrics.

Numbers. We conduct independent assessments for integers (e.g. 1887), decimals (e.g. 154.32), and large numbers (e.g. 200 billion). Large numbers have the format "integer/decimal million/billion/trillion". We create near-exhaustive candidate sets of valid number translations and check if the translation matches any candidate.

Physical Units. We build near-exhaustive candidate sets for evaluating the translations of diverse units including those related to weight, length, time, or temperature *inter alia* (e.g. inches). Translations are evaluated by string matching.

Emojis, Names, and Web Terms. Via string matching we check whether the translated text retains the same property instantiation found in the

source text. Candidate sets for these tests are thus considered to be exhaustive.

Currencies. We consider currencies appearing in the ISO code format (e.g. EUR). Near exhaustive candidate sets are built allowing translations into the same ISO code, variations of the currency name or its symbol (e.g. for En→Es: EUR/euro/euros/€), then a string matching pass-fail detection is employed.

Idioms. Idiomatic expressions pose significant challenges for MT systems due to their non-literal nature and potential large sequence length. We use idioms as a test bench for the use of contrastive candidate pairs (incorrect literal translation candidate vs. correct meaning translation) and semantic similarity detection procedure.

8 Models Comparison

In this section, we introduce the tested models and present results obtained via standard metrics as well as our proposed framework.

8.1 Experimental Setup

We test widely-used open-source MT models, as well as a commercial system. We aim to select models that perform very strongly, while also differing in some important aspects (e.g. bilingual vs. multilingual).

In the multilingual domain, we experiment with the 600M and the 3.3B parameters models of No Language Left Behind project (NLLB) (Team et al., 2022), and the Many-to-Many (MLM-100) family of Multilingual models (Fan et al., 2021) (418M and 1.2B parameters models). Additionally, we evaluate the WMT21: multilingual (7 En→X directions) 4.7B dense model (Tran et al., 2021), part of Meta's WMT-21 News Translation task participation (Barrault et al., 2021). We also assess OPUS-MT (Tiedemann and Thottingal, 2020) En→Es and En→De bilingual models trained on OPUS dataset (Tiedemann, 2012). Lastly, we included results from an anonymous commercial system.

Besides our proposed metrics, we also evaluate the models on FLORES-200 (Team et al., 2022) in En→De, En→Es, and En→Ja via string-based metrics spBLEU[10] (Papineni et al., 2002) and ChrF[11] (Popović, 2015) as implemented in SACREBLEU (Post, 2018), as well as the neural-based metric COMET-22[12] (Rei et al., 2020).

[10] SACREBLEU signature: `nrefs:1|case:mixed|eff:no|tok:flores101|smooth:exp|version:2.3.1`

[11] SACREBLEU signature: `nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.3.1`

Model	Source Sentence	Translation
OPUS MT (Bil)	The article I read on `www.scientificjournal.org` was very informative.	El artículo que leí en `www.cientificojournal.org` fue muy informativo.
Commercial system	... our town's population was counted as 12,577.	... población de nuestra ciudad se contabilizó en 12,577.

Table 4: Examples flagged as failed translations.

8.2 General Translation Accuracy

We first measure general translation performance across language pairs for standard reference-based metrics (Table 3). The commercial system performs best across the board, followed by the WMT21 model. In the following sections, we dive deeper into the different capabilities.

8.3 Behavioral Tests Results

As an illustrative example, macro pass rate confidence intervals across property types and models for the En→De direction are presented in Figure 5. The complete results can be found in Appendix E.

Commercial system is most consistent across properties. This is especially true for emoji translations, where open-source models lack most emojis in their vocabulary. However, it is noteworthy that its performance is subpar in the context of En→Es integers and En→Ja large numbers. After manual inspection (see examples in Table 4), we attribute the lower integers translation performance to the fact that it uses the comma as the thousands separator. Note that this behavior can be acceptable depending on the country; behavioral tests must be designed to reflect the intended behavior.

Bilingual models struggle with web terms. Although the multilingual models mostly manage to preserve web terms without alteration, both tested bilingual models (for En→Es and En→De) underperform in that property (see Figure 5 and Figure 6 top). Most fail cases contain Spanish words inside the translated web terms (Table 4). We hypothesize that this occurs because they are trained to exclusively translate into Spanish, which consequently hinders their ability to generate content in other

[12] `Unbabel/wmt22-comet-da`

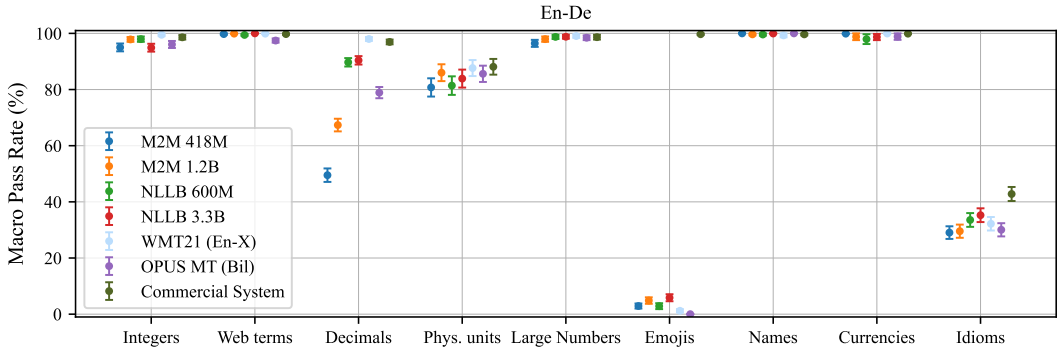


Figure 5: En→De macro pass rates and confidence intervals across tested systems.

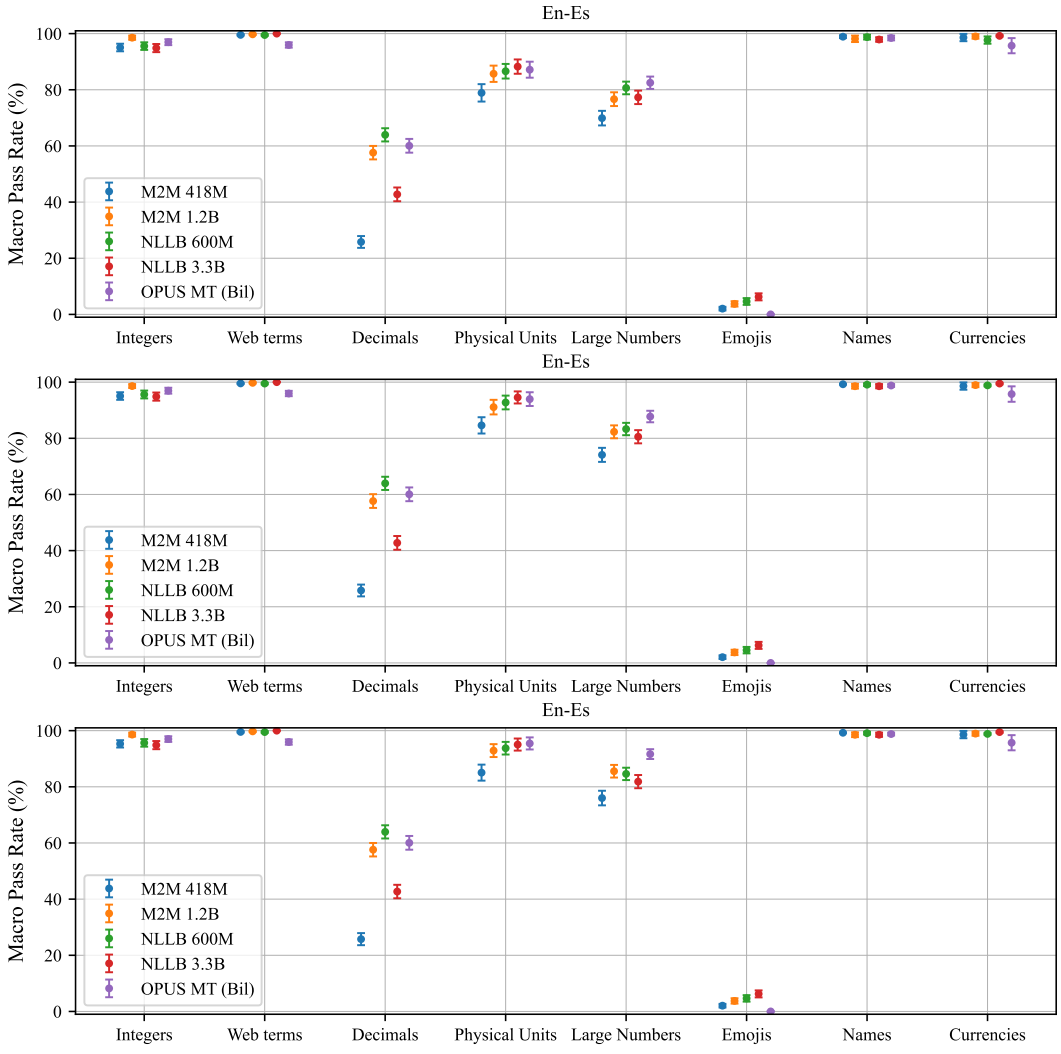


Figure 6: From top to bottom, En→Es Confidence Intervals after each annotation iteration (see §9).

languages, and is therefore an intrinsic limitation of bilingual models.

Scaling models help increase capabilities. In most of the settings, scaling the model of the same family shows increased performance, for instance, physical units and idioms in Figure 5. However, there are some counter-examples, like in the case of En→Ja decimals and integers tests.

WMT21 is the strongest open-source model. The WMT21 model consistently exhibits superior performance compared to other open-source models in both En→De and En→Ja tests. In Table 5 we show how paired bootstrap enables model comparison, revealing that WMT21 outperforms other

Model A	Model B	Winner	p-value
M2M 418M	M2M 1.2B	M2M 1.2B	0.0
M2M 418M	NLLB 600M	NLLB 600M	0.0
M2M 418M	NLLB 3.3B	M2M 418M	0.476
M2M 418M	WMT21 (En-X)	WMT21 (En-X)	0.0
M2M 418M	OPUS MT (Bil)	OPUS MT (Bil)	0.106
M2M 418M	Commercial system	Commercial system	0.0
M2M 1.2B	NLLB 600M	NLLB 600M	0.461
M2M 1.2B	NLLB 3.3B	M2M 1.2B	0.0
M2M 1.2B	WMT21 (En-X)	WMT21 (En-X)	0.001
M2M 1.2B	OPUS MT (Bil)	M2M 1.2B	0.004
M2M 1.2B	Commercial system	Commercial system	0.102
NLLB 600M	NLLB 3.3B	NLLB 600M	0.0
NLLB 600M	WMT21 (En-X)	WMT21 (En-X)	0.003
NLLB 600M	OPUS MT (Bil)	NLLB 600M	0.005
NLLB 600M	Commercial system	Commercial system	0.142
NLLB 3.3B	WMT21 (En-X)	WMT21 (En-X)	0.0
NLLB 3.3B	OPUS MT (Bil)	OPUS MT (Bil)	0.118
NLLB 3.3B	Commercial system	Commercial system	0.0
WMT21 (En-X)	OPUS MT (Bil)	WMT21 (En-X)	0.0
WMT21 (En-X)	Commercial system	WMT21 (En-X)	0.024
OPUS MT (Bil)	Commercial system	Commercial system	0.0

Table 5: Paired Bootstrap En→De Integers test results. We make a 95% statistically significant conclusion that the WMT21 system is better than the rest of the models.

models in the integers En→De test.

Idioms. Results for the Idioms property test are presented in Appendix D. The ability to translate idioms is generally low (i.e. overly literal), in accordance with recent findings (Dankers et al., 2022). It is worth noting that results are similar in the three language directions, with the commercial system and NLLB 3.3B showing comparable performance.

9 Reliability of the Proposed Approach

To assess the reliability of the proposed approach, in this section we analyze the robustness of source sentence generation and pass-fail detection.

9.1 Analysis of Source Sentence Generation

One potential concern with the proposed method is whether the generated source sentences are diverse enough and do not become repetitive after a few rounds of generation.[13] A standard method for quantifying the diversity in a corpus is *distinct n-grams* (Li et al., 2016), which computes the ratio of unique n-grams to the total number of n-grams present. In our case, we are interested in assessing the diversity of each generated source sentence compared to the previous generations. To that end, we propose a metric to measure this aspect. Given the set of unique n-grams generated up to sentence $\mathbf{x}_t$ ($\mathcal{G}^n_{\mathbf{x}_{<t}}$), we measure the proportion of unique

[13]The *naturalness* of outputs, another potential concern, has been extensively dealt with elsewhere (Ouyang et al., 2022).

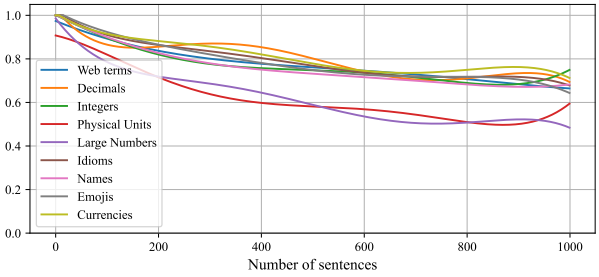


Figure 7: 3-gram diversity scores ($\text{div}_3(\mathbf{x}_t)$) along generation steps across different properties.

Property	Sentences kept	Unique values
Web Terms	79.3%	92.5%
Decimals	74.1%	76.9%
Integers	62.1%	39.3%
Physical Units	83.3%	15.9%
Large Numbers	66.1%	37.1%
Idioms	83.8%	69.0%
Names	86.1%	17.9%
Emojis	88.5%	29.7%
Currencies	66.8%	5.2%

Table 6: Percentage of source sentences that pass filtering, and percentage of filtered sentences that introduce a new property value.

n-grams in each newly generated sentence ($\mathcal{G}^n_{\mathbf{x}_t}$) that are not present in $\mathcal{G}^n_{\mathbf{x}_{<t}}$:

$$\text{div}_n(\mathbf{x}_t) = \frac{\mathcal{G}^n_{\mathbf{x}_t} \setminus \mathcal{G}^n_{\mathbf{x}_{<t}}}{\mathcal{G}^n_{\mathbf{x}_t}} \quad (6)$$

Figure 7 shows 3-gram diversity along 1000 generated sentences after fitting a polynomial regression. We observe that the diversity drop is mild even after 500 sentences, where for most of the tests, 60% of newly generated 3-grams are novel.

Furthermore, we observe that the sentence generator produces sentences that comply with instructions, indicated by the high proportion of the original sentences that pass filtering. In the majority of cases, over 70% of the LLM-generated sentences successfully pass the filtering steps outlined in §3, as seen in Table 6 (middle column). The right column shows the percentage of unique values, which naturally vary strongly depending on the property.

9.2 Analysis of Pass-Fail Detection

The reliability of the proposed pass-fail detection depends mainly on whether candidate sets are (1) complete and (2) do not contain wrong candidates.

We analyze this by sampling 100 random test cases that were marked as *pass* (positives), and an-

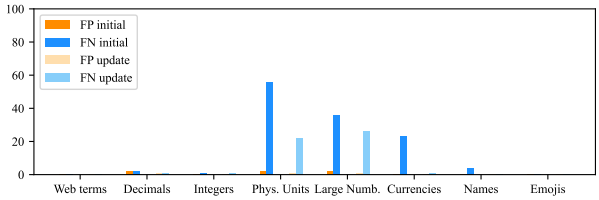


Figure 8: Error rates detected in two rounds of annotations on En→Es.

other 100 examples marked as *fail* (negatives). We manually annotate whether test results were correct or incorrect. Figure 8 shows false positives and false negatives (FP initial and FN initial). We observe that while for most properties these were low, for some test cases (namely physical units, large numbers, currencies) there were a significant number of FNs, which would lead to underestimated pass rates. We argue that erring on the side of FNs is generally preferable, because it prevents us from overestimating the strength of models, and because it would trigger a debugging effort which would quickly surface issues stemming from FNs.

To obtain more accurate pass-rates for all properties, we can manually remove candidates causing a FP and add missing candidates producing a FN. We do this for the test cases analyzed above, and then draw another random sample from both *pass* and *fail* categories. Figure 8 shows that the updated FPs and FNs are now negligible.

While in our experience, human intervention as outlined above is only a minor effort, the issue remains as to whether systems can be compared to one another without the need for human intervention, even in the presence of existing FNs. To understand this better, we plot macro pass rates with confidence intervals across annotation iterations in Figure 6. As expected, for physical units, large numbers, and currencies, pass rates move upwards. However, the effect is general across models, suggesting that relative ordering between models can be reasonably approximated in the initial attempt, i.e. without human intervention.

In addition, we assess the pass-fail detection of idioms. Given that the decision is made via semantic similarity for contrasting pairs, addressing issues in the candidate sets is more challenging. Consequently, we conducted a single evaluation iteration with 100 *pass*/*fail* examples, respectively, on two language pairs. For En→De, we observed 59 FPs / 16 FNs; En→Es had 50 FPs / 11 FNs. We hypothesize high FPs are caused by idiom and its figurative meaning being present within the source sentence, interfering with the n-grams comparison. We leave further investigation for future research.

10 Related Work

Recent works have applied behavioral testing for evaluating machine translation systems. Wang et al. (2021) designed tests for numerical translation capabilities by relying on fixed templates for source sentence generation. Raunak et al. (2022) proposed SALTED, a set of manually designed error detectors that are applied to millions of sentences from standard datasets. Beyond behavioral testing, a large number of *challenge sets* have been developed for machine translation (Popović and Castilho, 2019). Although useful, most of these evaluation tools require major human efforts for creation, evaluation, or expanding to other languages. Although there have been attempts to automatize the creation of behavioral tests (Yang et al., 2022), this has been limited to simple NLP tasks.

Our work also relates to the use of LLMs as evaluators for Machine Translation systems (Kocmi and Federmann, 2023), as well as for text generation in a broader sense (Liu et al., 2023; Xu et al., 2023), which extend the growing body of research on multi-dimensional text generation evaluation (Zhong et al., 2022; Yuan et al., 2021).

Behavioral testing aims to evaluate the behavior of systems under realistic conditions, contrasting it from the literature on adversarial data generation (Belinkov and Bisk, 2018; Zhang et al., 2021).

11 Conclusions

In this work, we have presented a method that automates the creation of behavioral tests to perform fine-grained evaluation of MT systems capabilities. We use Large Language Models to generate source sentences composed of fragments of specific language properties (integers, web terms, etc.), as well as translations of these properties. For property types formed by multiple words, we further extend the proposed method into a contrastive setting and show its usefulness in evaluating idiomatic expressions. To the best of our knowledge, our research represents the first attempt to develop MT behavioral tests by leveraging LLMs. Finally, we apply the proposed framework to evaluate open-source models on three language pairs.

References

Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors. 2021. *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics, Online.

B. Beizer and J. Wiley. 1996. Black box testing: Techniques for functional testing of software and systems. *IEEE Software*, 13(5):98–.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

B. Efron. 1979. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović and Sheila Castilho. 2019. Challenge test sets for MT evaluation. In *Proceedings of Machine Translation Summit XVII: Tutorial Abstracts*, Dublin, Ireland. European Association for Machine Translation.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Vikas Raunak, Matt Post, and Arul Menezes. 2022. SALTED: A framework for SAlient long-tail translation error detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5163–5179, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on*

Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. 2021. Facebook AI's WMT21 news translation task submission. In *Proceedings of the Sixth Conference on Machine Translation*, pages 205–215, Online. Association for Computational Linguistics.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Jun Wang, Chang Xu, Francisco Guzmán, Ahmed El-Kishky, Benjamin Rubinstein, and Trevor Cohn. 2021. As easy as 1, 2, 3: Behavioural testing of NMT systems for numerical translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4711–4717, Online. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Yang Wang, and Lei Li. 2023. Instructscore: Towards explainable text generation evaluation with automatic feedback.

Guanqun Yang, Mirazul Haque, Qiaochu Song, Wei Yang, and Xueqing Liu. 2022. TestAug: A framework for augmenting capability-based NLP tests. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3480–3495, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Xinze Zhang, Junzhe Zhang, Zhenhua Chen, and Kun He. 2021. Crafting adversarial examples for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1967–1977, Online. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Limitations

While the proposed evaluation framework seeks to address a broad spectrum of languages, the experiments conducted in this study are limited to three language pairs. Due to its reliance on the capacity of LLMs to produce high-quality candidate translations, we cannot guarantee accurate results when applied to language pairs involving a low-resource language using current LLMs. Moreover, the method is designed to work only on properties that appear as a continuous chunk of text in both source and target languages and are not scattered across a sentence.

B Example of Demonstrations for Exhaustive Candidate Set Generation

```
You are a {src_lang}-{tgt_lang} translator. Given a {property}, write as many valid
{target_lang} translations as you can. Use "|" to separate between valid translations.
Write "NA" if unable to accomplish the task.

EUR
€|EUR|Euro

GBP
£|GBP|Pfund|Pfund Sterling|britisches Pfund|Pound Sterling

USD
$|USD|Dollar|US Dollar|amerikanischer Dollar|amerikanische Dollar|US-Dollar

{Source property}
```

Figure 9: General template of the prompt used for generating a set of candidate translations.

C Example of Demonstrations for Contrastive Candidate Pairs Generation

Foil:

```
You are an {src_lang}-{tgt_lang} literal translator. Given a sequence of words,
you have to write only a literal translation. Use "|" to separate alternatives.
Write "NA" if unable to accomplish the task.

break a leg
brich dir ein Bein|breche dir ein Bein|breche dein Bein|breche dir dein Bein

hit the ground running
im Laufen hinfallen|beim Laufen hinfallen|beim Laufen auf den Boden knallen|beim Laufen
auf den Boden fallen

put all your eggs in one basket
alle Eier in einen Korb tun|alle Eier in einen Korb setzen|alle Eier in einen Korb legen

{Source property}
```

Correct:

```
You are an {src_lang_name}-{tgt_lang_name} translator of idiomatic expressions. Given an
idiomatic expression, you have to write the translation of the figurative meaning of the
idiomatic expression. Use "|" to separate alternatives. Write "NA" if unable to accomplish the task.

She told him to “break a leg” just before he went up on stage.
figurative translation of: break a leg
viel Glück|alles Gute|viel Erfolg|du schaffst das|Sie schaffen das

He hit the ground running, so his employer was really happy.
figurative translation of: hit the ground running
voller Begeisterung angehen|enthusiastisch angehen|hart und erfolgreich arbeiten

{Source property}
```

Figure 10: Prompt used for generating contrastive candidate pairs for the case of idioms. For the literal translation (foil) we prompt ChatGPT with the idiom in isolation. Conversely, in order to facilitate the ‘understanding’ of the idiom’s figurative connotation, for generating correct candidates we present it within the full sentence.

D Idioms Test Results

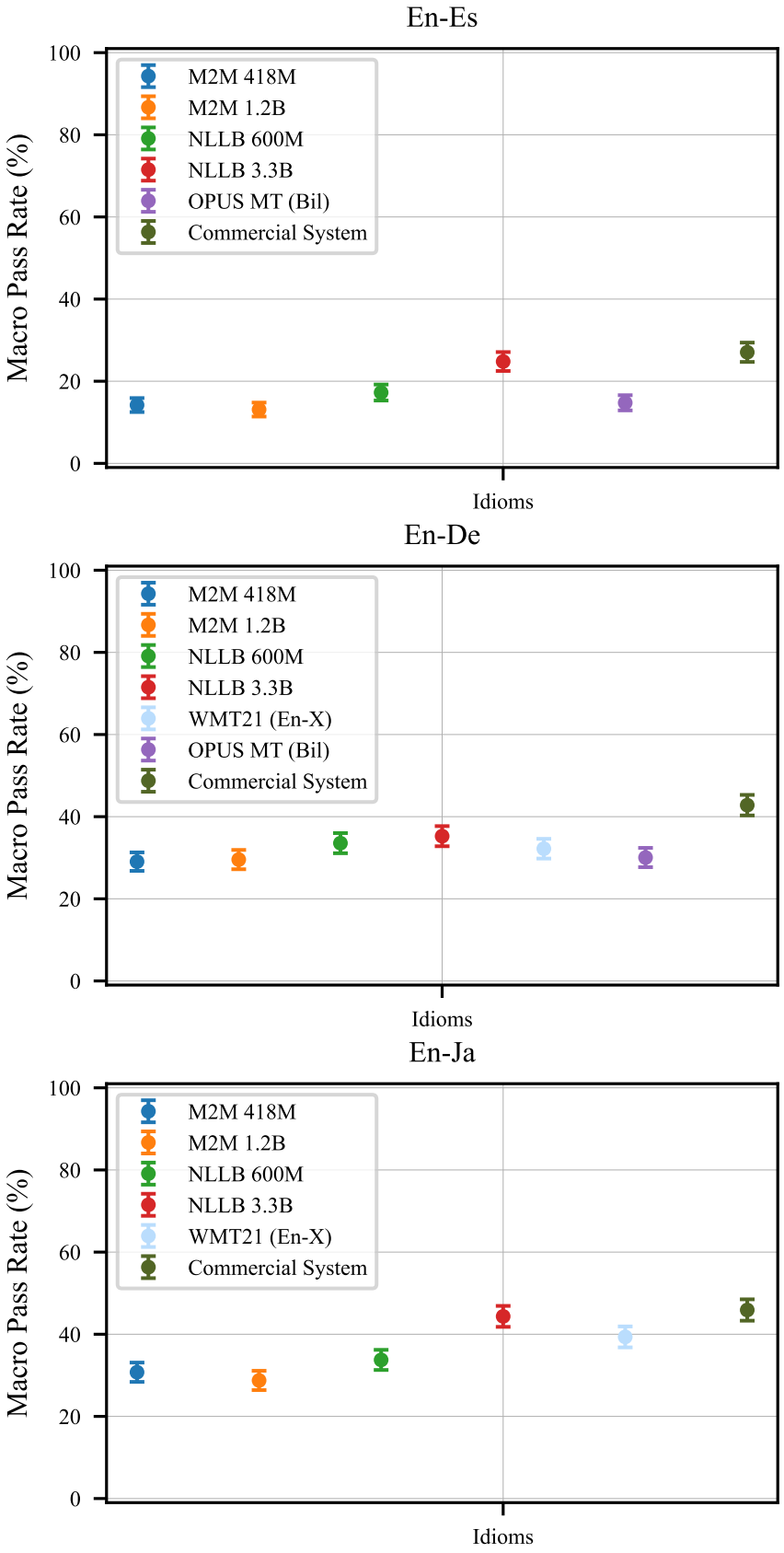


Figure 11: Idioms Test Confidence Intervals across language pairs.

E Pass Rate Confidence Intervals

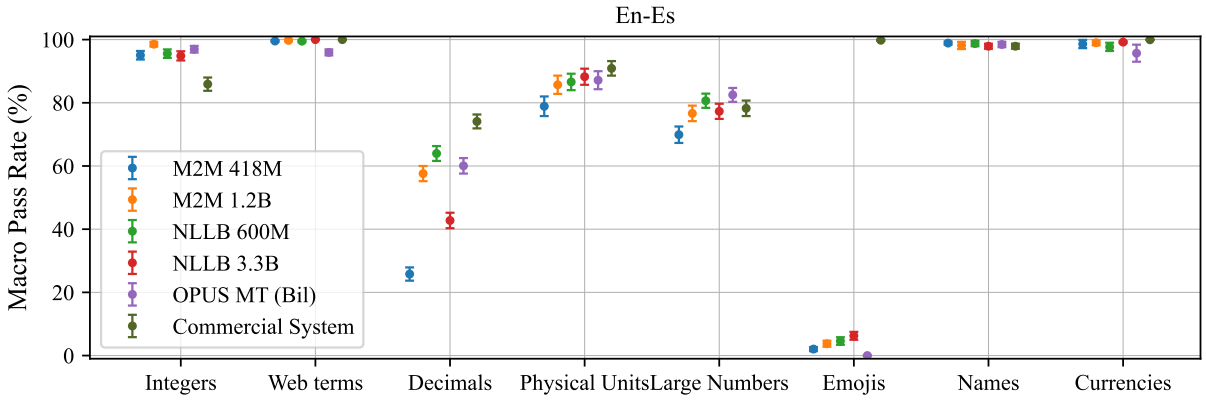


Figure 12: Macro pass rate confidence intervals for En→Es tests.

Property	Macro Pass Rate (%)					
	M2M 418M	**M2M 1.2B**	**NLLB 600M**	**NLLB 3.3B**	**OPUS MT (Bil)**	**Commercial system**
Web terms	[0.993, 0.998]	[0.995, 1.0]	[0.992, 0.998]	[1.0, 1.0]	[0.95, 0.97]	[1.0, 1.0]
Decimals	[0.236, 0.278]	[0.551, 0.602]	[0.615, 0.665]	[0.401, 0.451]	[0.576, 0.625]	[0.719, 0.764]
Integers	[0.939, 0.966]	[0.979, 0.993]	[0.943, 0.969]	[0.935, 0.963]	[0.959, 0.98]	[0.838, 0.88]
Physical Units	[0.824, 0.879]	[0.905, 0.952]	[0.915, 0.96]	[0.93, 0.972]	[0.933, 0.975]	[0.952, 0.987]
Large Numbers	[0.736, 0.787]	[0.834, 0.878]	[0.824, 0.868]	[0.795, 0.842]	[0.898, 0.935]	[0.863, 0.907]
Emojis	[0.014, 0.027]	[0.027, 0.048]	[0.035, 0.058]	[0.05, 0.075]	[0.0, 0.0]	[0.996, 1.0]
Names	[0.991, 0.994]	[0.977, 0.994]	[0.986, 0.996]	[0.978, 0.993]	[0.983, 0.993]	[0.978, 0.993]
Currencies	[0.973, 0.999]	[0.982, 0.997]	[0.985, 0.992]	[0.992, 0.998]	[0.93, 0.985]	[0.999, 1.0]
Idioms	[0.125, 0.159]	[0.114, 0.148]	[0.153, 0.192]	[0.225, 0.271]	[0.129, 0.166]	[0.247, 0.294]

Figure 13: Macro pass rate confidence intervals for En→Es tests.

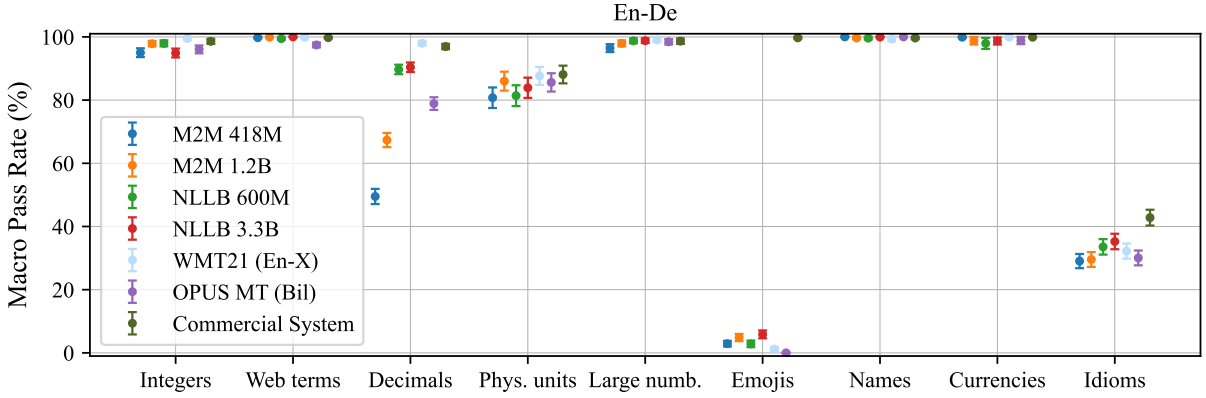


Figure 14: Macro pass rate confidence intervals for En→De tests.

Property	Macro Pass Rate (%)						
	M2M 418M	**M2M 1.2B**	**NLLB 600M**	**NLLB 3.3B**	**WMT21 (En-X)**	**OPUS MT (Bil)**	**Commercial system**
Web terms	[0.995, 1.0]	[0.998, 1.0]	[0.991, 0.998]	[1.0, 1.0]	[0.998, 1.0]	[0.967, 0.982]	[0.995, 1.0]
Decimals	[0.471, 0.519]	[0.651, 0.696]	[0.882, 0.912]	[0.889, 0.919]	[0.973, 0.987]	[0.769, 0.809]	[0.961, 0.978]
Integers	[0.936, 0.964]	[0.97, 0.987]	[0.97, 0.989]	[0.935, 0.963]	[0.99, 1.0]	[0.948, 0.973]	[0.978, 0.994]
Physical Units	[0.775, 0.84]	[0.83, 0.89]	[0.781, 0.847]	[0.807, 0.871]	[0.848, 0.905]	[0.827, 0.885]	[0.853, 0.909]
Large Numbers	[0.952, 0.977]	[0.97, 0.989]	[0.98, 0.995]	[0.981, 0.995]	[0.984, 0.997]	[0.976, 0.993]	[0.978, 0.995]
Emojis	[0.02, 0.038]	[0.037, 0.06]	[0.018, 0.039]	[0.046, 0.071]	[0.006, 0.018]	[0.0, 0.0]	[0.994, 1.0]
Names	[1.0, 1.0]	[0.993, 1.0]	[0.992, 1.0]	[0.999, 1.0]	[0.986, 1.0]	[1.0, 1.0]	[0.993, 1.0]
Currencies	[0.998, 1.0]	[0.976, 1.0]	[0.962, 0.997]	[0.976, 0.998]	[0.999, 1.0]	[0.977, 1.0]	[0.998, 1.0]
Idioms	[0.268, 0.313]	[0.272, 0.319]	[0.311, 0.36]	[0.328, 0.377]	[0.298, 0.346]	[0.277, 0.324]	[0.403, 0.453]

Figure 15: Macro pass rate confidence intervals for En→De tests.

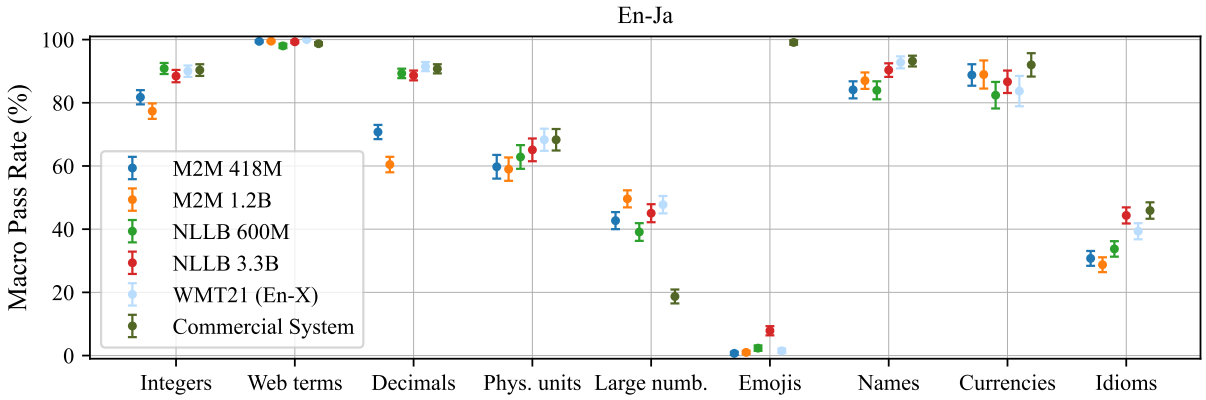


Figure 16: Macro pass rate confidence intervals for En→Ja tests.

Property	Macro Pass Rate (%)					
	M2M 418M	**M2M 1.2B**	**NLLB 600M**	**NLLB 3.3B**	**WMT21 (En-X)**	**Commercial system**
Web terms	[0.991, 0.998]	[0.992, 0.998]	[0.973, 0.987]	[0.989, 0.997]	[1.0, 1.0]	[0.982, 0.992]
Decimals	[0.685, 0.73]	[0.58, 0.629]	[0.878, 0.908]	[0.871, 0.902]	[0.9, 0.929]	[0.893, 0.922]
Integers	[0.795, 0.84]	[0.749, 0.798]	[0.891, 0.926]	[0.865, 0.904]	[0.882, 0.918]	[0.885, 0.922]
Physical Units	[0.56, 0.635]	[0.553, 0.627]	[0.591, 0.666]	[0.615, 0.687]	[0.648, 0.718]	[0.649, 0.717]
Large Numbers	[0.4, 0.454]	[0.469, 0.523]	[0.363, 0.419]	[0.422, 0.479]	[0.45, 0.505]	[0.165, 0.209]
Emojis	[0.002, 0.012]	[0.005, 0.015]	[0.015, 0.032]	[0.064, 0.093]	[0.008, 0.022]	[0.984, 0.998]
Names	[0.814, 0.868]	[0.844, 0.896]	[0.811, 0.868]	[0.882, 0.925]	[0.909, 0.947]	[0.915, 0.949]
Currencies	[0.854, 0.922]	[0.845, 0.934]	[0.782, 0.866]	[0.831, 0.902]	[0.789, 0.885]	[0.883, 0.957]
Idioms	[0.284, 0.331]	[0.264, 0.311]	[0.313, 0.362]	[0.418, 0.469]	[0.368, 0.419]	[0.433, 0.485]

Figure 17: Macro Pass Rate confidence intervals for En→Ja tests.

One Wide Feedforward is All You Need

Telmo Pessoa Pires[*†] **António V. Lopes** **Yannick Assogba** **Hendra Setiawan**[*]
Equall
telmo@equall.ai
Apple
{antoniovilarinholopes, yassogba, hendra}@apple.com

Abstract

The Transformer architecture has two main non-embedding components: Attention and the Feed Forward Network (FFN). Attention captures interdependencies between words regardless of their position, while the FFN non-linearly transforms each input token independently. In this work we explore the role of the FFN, and find that despite taking up a significant fraction of the model's parameters, it is highly redundant. Concretely, we are able to substantially reduce the number of parameters with only a modest drop in accuracy by removing the FFN on the decoder layers and sharing a single FFN across the encoder. Finally we scale this architecture back to its original size by increasing the hidden dimension of the shared FFN, achieving substantial gains in both accuracy and latency with respect to the original Transformer Big.

1 Introduction

The Transformer architecture (Vaswani et al., 2017) has become the de facto paradigm in many Natural Language Processing (NLP) tasks, including Machine Translation (MT). Several studies have shown that Transformers exhibit impressive scaling-law properties (Gordon et al., 2021; Bansal et al., 2022; Ghorbani et al., 2022), wherein increasing the number of model parameters leads to further accuracy gains. In parallel with this architecture's impressive scaling of the numbers of parameters (Chowdhery et al., 2022), there is a growing trend towards reducing model footprints for real-world deployment, to satisfy practical constraints like latency requirements as well as memory and disk space limitations. In turn, researchers are actively exploring parameter sharing (Ge et al., 2022; Takase and Kiyono, 2023; Lou et al., 2022), reducing the dimensionality of Transformer components, and pruning components like attention heads (Voita et al., 2019; Michel et al., 2019).

Although the role of attention in learning pairwise dependencies between tokens is relatively well understood (Voita et al., 2019; Clark et al., 2019; Vig and Belinkov, 2019), the role of the Feed Forward Network (FFN) remains under-explored. Recently, Geva et al. (2021) established a connection between the FFN and attention by positing that the FFN corresponds to learnable *key-value* pairs where the weights of the first layer of the FFN corresponds to the *keys* and those of the second to the *values*. They find that the keys are able to capture salient textual patterns at each layer, and they notice that the classes of patterns tend to overlap between neighboring layers, indicating redundancy in the representation.

This observation motivates our work, where we revisit the conventional practice of allocating an individual FFN per layer. We investigate the effect of sharing and dropping the FFN across different layers on MT models. We conduct thorough experiments with different configurations of the Transformer, across different language pairs, including a low resource language pair and multilingual. In addition, we investigate the effect of the FFN in a decoder-only Transformer-based model. We find that a considerable level of redundancy exists between the encoder and decoder FFNs. As a result, we are able to eliminate the decoder FFN and share a single FFN across the encoder without significantly compromising the model's accuracy. This step leads not only to significant parameter savings but also opens up opportunities for further improvements. We also suggest using wider FFNs in the encoder while dropping the decoder's FFN, which results in a model with a similar size, but improved accuracy and reduced latency.

Finally we conduct a fine-grained analysis of the representational similarity between the original model, using one independent FFN per layer,

[*]Equal contribution.
[†]Work conducted while at Apple.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 1031–1044
December 6–7, 2023.

and various models with shared FFNs. Our results reveal that both model accuracy and the internal representation of Transformer blocks remain stable when sharing the FFN.

2 Background and Methodology

2.1 Transformer

The Transformer architecture has two main components: attention and the FFN, which are connected via a residual connection (He et al., 2016) and layer normalization (Ba et al., 2016). In an encoder-decoder model, there are two types of attention: self-attention and cross-attention. Self-attention is used in both the encoder and the decoder, allowing the model to focus on relevant information within the same sequence. Cross-attention is exclusive to the decoder and allows it to attend to the encoder's output. Attention takes as input a set of *queries*, *keys* and *values*, projected using four $\mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ matrices (one for the queries, keys, values, and final output) where d_{model} is the model's hidden dimension. It then applies the SOFTMAX function to allow it to focus on the most relevant values.

The FFN is applied after attention on both the encoder and the decoder and consists of the following 2-layer linear transformation:

$$\text{FFN}(\boldsymbol{x}) = \max(0, \boldsymbol{x}\mathbf{W}_1 + b_1)\mathbf{W}_2 + b_2, \quad (1)$$

where a RELU non-linearity is applied to the transformation of the input sequence ($\boldsymbol{x}$). At each layer, the FFN is parameterized with two matrices, $\mathbf{W}_1 \in \mathbb{R}^{d_{\text{model}} \times d_{\text{ff}}}$ and $\mathbf{W}_2 \in \mathbb{R}^{d_{\text{ff}} \times d_{\text{model}}}$ where d_{ff} is the *FFN dimension* and is usually set to $4 \times d_{\text{model}}$ (Vaswani et al., 2017).

Recent work has drawn a significant link between attention and the FFN (Geva et al., 2021), wherein $\mathbf{W}_1$ and $\mathbf{W}_2$ assume roles akin to the *keys* and *values* to an unnormalized attention where the input ($\boldsymbol{x}$) acts as the *query*. Unlike regular attention, the FFN employs a RELU, which allows *multiple* keys to significantly contribute to the final output (Geva et al., 2021). Additionally, these keys correspond to an inventory of salient patterns that are learned from the training data. Geva et al. (2021) suggest that at the lower layers the FFN learns shallow syntactic patterns and progressively learns deep semantic patterns on the deeper layers. Moreover, the authors find that there's a substantial overlap between patterns captured by adjacent layers, indicating that there are redundancies in the FFNs and suggesting a better allocation of these parameters might be beneficial for performance.

2.2 Sharing and Widening the FFN

The vanilla Transformer allocates one FFN for each layer of the encoder and decoder, i.e. FFN_i^{enc} or FFN_i^{dec}, respectively. Excluding embedding parameters, these FFNs occupy around two thirds of the parameter budget, while attention occupies the remaining third[1]. Earlier work found that constraining the parameterization of the decoder FFNs causes no degradation in accuracy (Ge et al., 2022). In this work, we share the parameters of the FFN across layers and/or across the encoder and decoder to minimize redundancy between FFNs.

Let N_{enc}, N_{dec} be the numbers of encoder and decoder layers, respectively. We consider multiple configurations for parameter sharing as follows:

- One $\textbf{FFN}_{all}^{enc}$ for the whole encoder:

  $$\text{FFN}_i^{enc}(\cdot) \overset{\text{tied}}{=} \text{FFN}_{all}^{enc}(\cdot), \forall i : 1 \leq i \leq N_{enc}$$

- One $\textbf{FFN}_{all}^{dec}$ for the whole decoder:

  $$\text{FFN}_j^{dec}(\cdot) \overset{\text{tied}}{=} \text{FFN}_{all}^{dec}(\cdot), \forall j : 1 \leq j \leq N_{dec}$$

- One $\textbf{FFN}_{all}^{encdec}$ for both the encoder and the decoder:

  $$\text{FFN}_i^{enc}(\cdot) \overset{\text{tied}}{=} \text{FFN}_j^{dec}(\cdot) \overset{\text{tied}}{=} \text{FFN}_{all}^{encdec}(\cdot),$$
  $$\forall i, j : 1 \leq i \leq N_{enc}, 1 \leq j \leq N_{dec}$$

Additionally, we explore modifying the dimension of the shared FFN, which we denote as $d_{\text{ff}'}$. Setting $d_{\text{ff}'} > d_{\text{ff}}$ widens the shared FFN while $d_{\text{ff}'} < d_{\text{ff}}$ narrows it. We also consider the extreme cases of setting $d_{\text{ff}'}$ to 0 or to $(N_{enc} + N_{dec}) \times d_{\text{ff}}$ (and beyond). Setting $d_{\text{ff}'} = 0$ is equivalent to dropping the FFN[2] while setting $d_{\text{ff}'} = (N_{enc} + N_{dec}) \times d_{\text{ff}}$ is akin to sharing the concatenation of all individual FFNs.

Sharing the FFNs directly affects the number of parameters and, to a certain extent, latency. For instance, sharing FFN_{all}^{enc} for the whole encoder reduces the number of parameters by $(N_{enc} - 1) \times 2 \times d_{\text{model}} \times d'_{\text{ff}}$[3]; whereas removing the FFN on the

[1] Ignoring layer normalization, there are $4 \times d_{\text{model}} \times d_{\text{model}}$ parameters for attention vs $2 \times d_{\text{model}} \times d_{\text{ff}} = 8 \times d_{\text{model}} \times d_{\text{model}}$ parameters for the FFN, assuming $d_{\text{ff}} = 4 \times d_{\text{model}}$.

[2] In our experiments without the FFN (i.e., $d_{\text{ff}'} = 0$) we remove the residual connection and layer normalization associated with it, as they become redundant.

[3] Plus the layer normalization parameters, which we are ignoring for simplicity.

decoder, i.e., setting $d_{\text{ff}'} = 0$ for FFN^{dec}_{all}, reduces the parameters by $(N_{dec}) \times 2 \times d_{\text{model}} \times d'_{\text{ff}}$ and reduces the amount of computation to be done. This is particularly important during inference since the forward pass of the decoder is autoregressive, and changing the decoder's FFN dimension has a higher latency impact than on the encoder.

Since different configurations have different impacts, we analyse the trade-off between model size, latency, and accuracy: (i) How many parameters can be shared/pruned with negligible (if any) accuracy degradation? (ii) Are the encoder and decoder FFNs affected similarly? (iii) Keeping the same model size, can the FFN parameters be allocated more efficiently?

We propose a novel configuration, which we call the *One Wide FFN* model, consisting of a single shared wide FFN on the encoder and no FFN on the decoder. To keep the number of parameters the same as in the baseline, we increase the shared FFN dimension accordingly: FFN^{enc}_{all} with $d_{\text{ff}'} = (N_{enc} + N_{dec}) \times d_{\text{ff}}$.

For completeness, we include similar experiments on the attention mechanism in Appendix B. These experiments show that, contrary to the FFN, individual layer-specific attention weights are more important and not as redundant, as sharing the attention leads to significant accuracy drops.

2.3 Representational Similarity

Besides investigating the impact on accuracy, we study the similarity between different models in terms of their *internal representations* and the *semantic space* they produce.

We use Linear Centered Kernel Alignment (CKA, Kornblith et al., 2019) to measure the similarity between the internal representations of different models. CKA uses inner products to estimate how similar the kernel matrices of two different representations are, and is based on the Hilbert-Schmidt Independence Criterion (HSIC, Gretton et al., 2005), a statistical measure of independence of two random variables. Linear CKA uses the dot product as a kernel and can be written as:

$$\text{CKA}(\mathbf{A}, \mathbf{B}) = \frac{||\mathbf{A}\mathbf{B}^{\text{T}}||^2_{\text{F}}}{||\mathbf{A}^{\text{T}}\mathbf{A}||_{\text{F}}||\mathbf{B}^{\text{T}}\mathbf{B}||_{\text{F}}},$$

where $||\cdot||_{\text{F}}$ is the Frobenius norm while $\mathbf{A}$ and $\mathbf{B}$ are mean-centered (i.e., we subtract the mean) feature matrices of the layers under comparison, computed on the same dataset. Both matrices are $n \times d$, where n is the number of sentences in the dataset and d is the output dimension of the component, and are obtained by averaging the activation of all tokens in each sentence[4]. The linear kernel is straightforward to compute and Kornblith et al., 2019 report strong empirical performance of linear CKA compared to other kernels and methods.

To measure the similarity between the semantic spaces of different models, we use Local Neighborhood Similarity (LNS, Boggust et al., 2022). Local neighborhood similarities have been previously been used in analyzing semantic shifts in word embeddings (Hamilton et al., 2016). The premise of LNS is that two semantic spaces are similar if a sentence has similar neighbors in the two spaces. The LNS of a sentence s between models 1 and 2 is defined as:

$$\text{LNS}(s) = Sim(k\text{-NN}_1(s),\ k\text{-NN}_2(s)),$$

where k-NN(s) is the set of k nearest neighbors of sentence s for a model and Sim is the intersection-over-union (Jaccard similarity) of the two sets of neighbors. For each pair of components (attention and FFN) in models 1 and 2 we compute the LNS of all sentences in the evaluation dataset and take the mean LNS as our layer similarity measure. The smaller the value of k the more local the neighborhoods we are comparing, and the more specific the retrieval task. We pick k to be small enough to visually inspect sentence neighborhoods if necessary. In our analysis, we use cosine distance as the distance metric between activations and set k to 5% of the dataset size ($\sim$ 100 sentences).

3 Experimental Setup

Data In our experiments, we show results on WMT22 English (EN) $\rightarrow$ German (DE) (296M pairs), which we obtained using the provided `mt-data` scripts[5], WMT16 EN $\rightarrow$ Romanian (RO) (610K pairs), and for the multilingual setup of Pires et al. (2023), consisting of 10 languages: German, English, Spanish, French, Italian, Japanese, Korean, Portuguese, Swahili, and Chinese. In our analysis, we mostly focus on WMT22 EN $\rightarrow$DE.

Following Schmidt et al. (2022), we use WMT'16 provided scripts to normalize the RO side. EN $\rightarrow$RO keeps diacritics for producing accurate translations. For more details refer to Schmidt et al.

[4]We use the source sentence and force decode the first reference to compute the encoder and decoder representations, respectively.

[5]https://www.statmt.org/wmt22/mtdata/

(2022). For the multilingual experiments, we replicated the setup of Pires et al. (2023), which includes all details, including data preprocessing and dataset sizes.

Metrics We compute BLEU[6] using `sacreBLEU`[7] version 2.3.1, with evaluation signatures `nrefs:1 | case:mixed | eff:no | tok:13a | smooth:exp` for BLEU, and `nrefs:1 | case:mixed | eff:no | tok:flores101 | smooth:exp` for SPBLEU. For our main results, we also report COMET using the `wmt20-comet-da` model and CHRF using the signature `nrefs:1 | case:mixed | eff:yes | nc:6 | nw:0 | space:no`.

Latency We report inference time in tokens/second (the higher, the better), averaged over 5 runs. For the multilingual models, we use the DE →EN test set. Our measurements were collected using a single NVIDIA V100 GPU on a single-threaded Intel(R) Xeon(R) Gold 6148 CPU @ 2.40GHz with batch size of 1 and beam size of 5, in order to realistically mimic the inference of a deployed model. For experiments with larger batch sizes, see Appendix D.

Tokenization For WMT22 EN →DE, we use SENTENCEPIECE (Kudo and Richardson, 2018), with a vocabulary size of 32K and a character coverage of 1.0, while for the multilingual experiments we use a vocabulary size of 250k and a character coverage of 0.9995. For WMT16 EN →RO we use byte-pair encoding (BPE, Sennrich et al., 2016) with $40,000$ merge operations.

Model Architectures We focus our analysis on the Transformer `Big` where $N_{enc} = N_{dec} = 6$, $d_{model} = 1024$, $d_{\text{ff}} = 4096$, and it has 16 attention heads. We also report results on Transformer `Base` ($N_{enc} = N_{dec} = 6$, $d_{model} = 512$, $d_{\text{ff}} = 2048$, and 8 attention heads), and a deep encoder shallow decoder (Kasai et al., 2021) Transformer Big with 12 encoder layers, and 2 decoder layers. For our decoder-only experiments, the model is identical to the Transformer Big, except that all 12 layers are on the decoder. Our decoder-only model is similar to a Transformer-based language model, particularly Prefix-LM (Raffel et al., 2020), where we apply a non-autoregressive mask on the source side and an autoregressive mask on the target. The source and target embeddings and the output projection matrix are shared in all models (Press and Wolf, 2017).

Hyperparameters All experiments are implemented using FAIRSEQ (Ott et al., 2019). Our optimizer is ADAM (Kingma and Ba, 2015) with a learning rate of 0.0007. We train for 80k, 80k, 150k steps on WMT22, WMT16, and multilingual, respectively, at which point the models had converged. We use 4000 warm-up steps, and an inverse square root learning rate scheduler (Vaswani et al., 2017). We use a dropout rate of 0.1 for WMT22, 0.3 for WMT16, and 0 for the multilingual experiments due to the abundance of data, following Pires et al. (2023). All models are trained using `fp16` (Ott et al., 2018).

Nomenclature In our experiments, we run a number of different configurations per model architecture that differ in the way the FFN is used, shared, or dropped, as well the size of the shared FFN ($d_{\text{ff}'}$). To facilitate our discussion, we introduce in Table 1 the nomenclature that will serve as reference for the rest of the text. Unless otherwise stated, the dimension of the shared FNN^{*}_{all}, i.e. $d_{\text{ff}'}$ is equal to the d_{ff} of the original model.

For decoder-only models, only `SharedDec` and `NoDec` configurations are defined. For conciseness, we drop the mention of FFN from the text when possible, i.e. `SharedEnc` instead of `SharedEncFFN`.

FFN Description	Encoder	Decoder
`SharedEnc`	FNN^{enc}_{all}	FNN^{dec}_{i}
`SharedDec`	FNN^{enc}_{i}	FNN^{dec}_{all}
`SharedEncSharedDec`	FNN^{enc}_{all}	FNN^{dec}_{all}
`SharedEncDec`	$\text{FNN}^{encdec}_{all}$	
`NoDec`	FNN^{enc}_{i}	`No-op`
`SharedEncNoDec`	FNN^{enc}_{all}	`No-op`

Table 1: Nomenclature used in our experiments. `No-op` indicates an identity function, which is equivalent to dropping the FFN.

Representational Similarity We use the WMT22 EN →DE evaluation set for both CKA and LNS analysis. We analyze encoder and decoder representations independently and present these metrics in a matrix heatmap plot showing pairwise similarity between layers. The diagonal of this matrix is the similarity of corresponding layers, i.e., layer i on both architectures. In order to

[6]For the multilingual experiments, we select the Flores101 tokenizer in `sacreBLEU`, so technically we report SPBLEU.

[7]https://github.com/mjpost/sacrebleu

facilitate an "apples-to-apples" comparison across models, we extract decoder representations by force decoding the (first) reference. We establish 2 crucial similarity scores: a *benchmark* on similarity for each of these metrics, where we train two additional models using the same architecture but with different random seeds; a similarity lower bound, where we compare the baseline Transformer Big with a randomly initialized (i.e., untrained) model with the same architecture. We present these bounds in Appendix C.

4 Experimental Results

4.1 Sharing FFNs

The results of various FFN sharing configurations are summarized in Table 2, including their impact on accuracy and model size (in millions of parameters and percentage). Sharing either the encoder (SharedEnc) or the decoder FFN (SharedDec) results in just a 0.2 to 0.3 BLEU point decrease, while reducing the parameter count by nearly 20%. Sharing the FFN on each side (ShareEncShareDec) leads to a more substantial degradation of 0.9 BLEU points, albeit reducing the parameter count by 37%, while sharing a single FFN on the encoder and decoder (ShareEncDec) results in a slightly higher degradation of 1.1 BLEU points. Nonetheless, these findings support the hypothesis that the FFN contains some degree of redundancy, as we expected a greater accuracy degradation given the substantial $(20-40\%)$ reduction in model size.

Architecture	BLEU	$\lvert\theta\rvert$	(%)
Transformer Big	35.6	228M	(100)
+ SharedEnc	35.4	186M	(82)
+ SharedDec	35.3	186M	(82)
+ SharedEncSharedDec	34.7	144M	(63)
+ SharedEncDec	34.5	136M	(59)

Table 2: sacreBLEU results on WMT 22 EN →DE for different FFN sharing configurations. $\lvert\theta\rvert$ is the number of parameters.

While we focus on sharing *one* FFN for all layers within a module, we compare with sharing multiple FFNs following Takase and Kiyono (2023) in Appendix A. We find that sharing one FFN is as accurate as sharing multiple FFNs within a module, while being more parameter-efficient.

4.2 Dropping FFNs

Table 3 summarizes the performance of models with no FFNs. Besides BLEU and number of parameters, we report the inference speed for each architecture. Dropping the FFN on the encoder (NoEnc) leads to a 0.9 BLEU point drop while reducing the parameter count by 22% and with minimal effect on inference speed. Dropping the FFN on the decoder (NoDec), on the other hand, causes a degradation of only 0.4 BLEU points while increasing the inference speed by 20%[8]. The highest latency reduction is obtained by removing the FFNs on both the encoder and the decoder (NoEncNoDec), but it comes with a significantly larger degradation of over 2 BLEU points.

Architecture	BLEU	Speed	$\lvert\theta\rvert$	(%)
Transformer Big	35.6	$111^{\pm 1.2}$	228M	(100)
+ NoEnc	34.7	$112^{\pm 1.0}$	178M	(78)
+ NoDec	35.2	$133^{\pm 0.9}$	178M	(78)
+ NoEncNoDec	33.5	$138^{\pm 1.9}$	127M	(56)
+ SharedEncNoDec	35.3	$136^{\pm 1.1}$	136M	(60)
+ NoEncSharedDec	33.9	$127^{\pm 1.0}$	136M	(60)

Table 3: sacreBLEU results on WMT 22 EN →DE for different FFN dropping configurations.

Combining sharing and dropping These results, together with those from Table 2, suggest that the encoder and decoder FFNs have different contributions: the decoder's are more redundant, corroborating previous work on FFNs parametrization (Ge et al., 2022). With this in mind, we experiment with one shared FFN on the encoder and dropping it on the decoder, reported as SharedEncNoDec in Table 3. As shown, with just approximately 60% of Transformer Big parameters we observe a 22% improvement in inference speed, at the cost of 0.3 BLEU point.

4.3 One Wide FFN Model

Previous sections describe models that share and/or drop FFNs, effectively reducing model size at some modest accuracy cost. In this section, we investigate whether we can regain the accuracy lost while preserving the parameter efficiency and the latency reduction. We focus on ShareEncNoDec model as

[8]The reason for this difference between NoEnc and NoDec is that the encoder output is computed in parallel, while the decoder operates in a step-by-step fashion.

	BLEU	CHRF	COMET	Speed	$\mid \theta \mid$ (%)
Transformer Big EN →DE	35.6	62.6	57.2	$110.8^{\pm 1.2}$	228M (100)
+ SharedEncNoDec FFN $d_{\text{ff}'} = 4,096$	35.3	62.1	56.1	$135.7^{\pm 1.1}$	135M (60)
+ SharedEncNoDec FFN $d_{\text{ff}'} = 24,576$	35.7	62.7	57.9	$138.2^{\pm 0.9}$	177M (80)
+ SharedEncNoDec FFN $d_{\text{ff}'} = 49,152$	**36.5**†	**63.2**†	**59.6**	$137.5^{\pm 1.6}$	228M (100)
+ SharedEncNoDec FFN $d_{\text{ff}'} = 98,304$	36.4†	**63.2**†	59.0	$134.5^{\pm 1.6}$	328M (145)

Table 4: Accuracy of One Wide FFN for Transformer Big EN →DE on WMT22. † implies the system is statistical significantly different at $p < 0.05$.

it provides a strong baseline with significant parameter savings and inference speedups.

We propose increasing the dimension of the shared FFN to match the number of parameters of the original (fully-parameterized) model, so as to avoid increasing the overhead of model storage. In particular, ShareEncNoDec saves around $(N_{enc} + N_{dec} - 1) \times 2 \times d_{model} \times d_{\text{ff}}$ parameters as there's one single shared FFN in the encoder. On the other hand, the Transformer Big has $(N_{enc} + N_{dec})$ FFNs. Thus, we match the size of the original model by setting the dimension of the shared FFN, $d_{\text{ff}'}$, to $(N_{enc} + N_{dec}) \times d_{\text{ff}}$.

Table 4 summarizes our results. It includes our proposed model, the *One Wide FFN* model ($d_{\text{ff}'} = 49,152$), as well as the baseline Transformer Big, and the corresponding ShareEncNoDec ($d_{\text{ff}'} = 4,096$). It also includes a wide model with $d_{\text{ff}'} = 24,576$, which uses the same number of parameters as NoDec, with $d_{\text{ff}'} = N_{enc} \times d_{\text{ff}}$. This model achieves an accuracy on par (or slightly above) the baseline Transformer Big with 20% fewer parameters and a significant inference speed-up.

Our proposed model with $d_{\text{ff}'} = 49,152$ goes beyond that, achieving a gain of 1.2 BLEU points over the vanilla ShareEncNoDec and 0.9 BLEU points over the Transformer Big. These gains remain consistent across CHRF and COMET. Furthermore, it has a similar inference speed as the ShareEncNoDec model. For completeness, we include a wider model with $d_{\text{ff}'} = 98,304$. Despite the extra capacity, this model does not provide any additional accuracy gains, which we suspect is due to the lack of data to train such a large model.

4.4 Analyzing Internal Representations

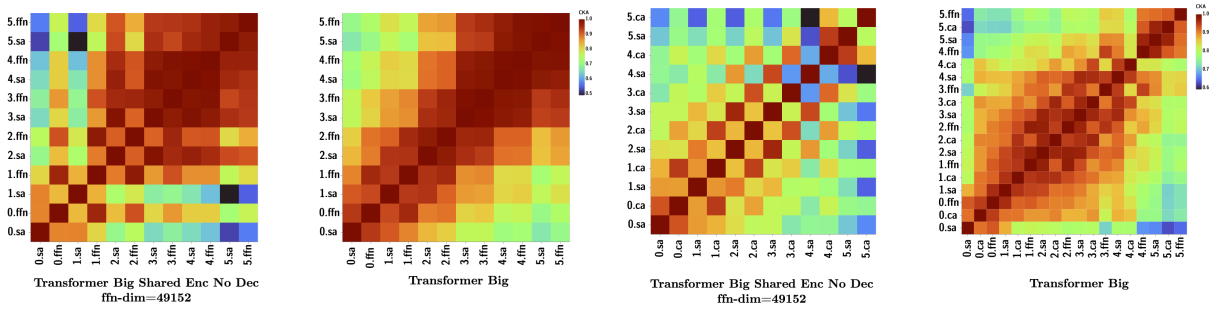
We now report a *post-hoc* analysis of the internal representations of the models introduced in preceding sections. Our objectives are twofold: 1) to ascertain whether the proposed models' internal representations exhibit a significant degree of similarity to those of the original base model; 2) to delve into the impact of the proposed methods on redundancy. We adopt the definition of redundancy of Dalvi et al. (2020), who *visually* inspect the similarity between adjacent modules within a model (high similarity entails high redundancy).

Architecture	Encoder		Decoder	
	CKA	LNS	CKA	LNS
Benchmark	100.0	100.0	100.0	100.0
SharedEnc	98.0	96.2	100.8	100.6
SharedDec	100.2	101.4	98.3	94.6
SharedEncSharedDec	98.9	97.2	99.5	95.4
SharedEncDec	97.6	94.4	98.4	93.5
NoEnc	90.0	70.5	101.0	96.8
NoDec	100.0	98.6	96.0	87.4
SharedEncNoDec	97.6	98.9	97.5	89.0
SharedEncNoDec$^{d'_{\text{ff}}=49152}$	97.0	83.2	94.0	82.9

Table 5: Similarity of the representations (%) of corresponding modules of different architectures vs. the Transformer Big for WMT22 EN →DE. These scores are normalized by comparing them to the CKA and LNS benchmark scores. For NoDec configurations we compare the final output of the Transformer layer as a whole as they have different modules than the baseline. The columns for shared and for dropped FFNs are highlighted in gray and blue respectively.

4.4.1 Similarity to Baseline

We ground the pairwise similarity metrics, by normalizing them against a benchmark. As mentioned in Section 3, we establish the *benchmark scores* by training two additional Transformer Big models, but using different random seeds. These models achieve similar accuracy as the baseline model (see Appendix C.1 for more details). The benchmark score is the similarity between the baseline and these models Because the benchmark is calculated by averaging similarity scores from different train-



(a) Encoder self similarity.

(b) Decoder self similarity.

Figure 1: CKA self similarity of encoder and decoder layers of the *One Wide Encoder* model vs. the Transformer Big baseline. We identify each component with a label: `index.name`. For example, `0.sa` refers to the self-attention on layer 0, while `4.ca` refers to the cross-attention on layer 4.

ing runs of our baseline, individual runs can have a normalized score above 100%.

Table 5 shows normalized similarity scores for several models. Under the Encoder columns we compare the encoder representations, and under the Decoder columns we compare decoder representations. Sharing FFNs leads to consistenly lower (normalized) similarity scores than models that do not share, both in terms of internal representation (CKA) and semantic spaces (LNS). As shown, although models that share FFNs have lower similarity scores compared to those that do not, the scores are still very close to 100%. Moreover, these decreases align with the drops in BLEU seen in Table 2, where the model with the lowest similarity score (`ShareEncDec`) is also the least accurate model. We observe a similar trend for models that drop the FFNs in the encoder or decoder, these models exhibit lower similarity scores with the respective component than models sharing them, as shown by `NoEnc` and `NoDec`. In addition, the former result again suggests the FFNs in the encoder are more important than in the decoder as the similarity shifts drastically compared to all other settings.

For completeness, we report on the last row the similarity scores for the One Wide FFN model, which is more accurate than the base model. The internal representations generated by that model diverge from those of the base model. Interestingly, we observe a larger drop in LNS scores than in CKA scores, indicating that the shift occurs mostly in semantic space, rather than the Euclidean space captured by CKA. For a detailed layer-wise similarity analysis that breaks out the aggregate analysis in Table 5 see Appendix C.2.

4.4.2 A Qualitative View of Redundancy

We now study into the impact of our One Wide FFN model on the redundancy of the internal representations. In addition to adopting their definition of redundancy, we also adopt Dalvi et al. (2020)'s method of computing self-similarity, namely looking at how the representations change as they go through each module (self-attention, FFN, or cross-attention) of the model. In particular, we use CKA to compute similarity between the output of different modules within the same model.

In Figure 1a, we show the CKA self-similarity matrices for the encoders of the One Wide FFN model and the `Transformer Big`. We do the same for the decoders in Figure 1b. These matrices show how similar each module of the network is to all other modules *within that network*. The diagonal of the matrix is the similarity between a module and itself and is always 1.

As shown, there is high similarity between adjacent modules of the `Transformer Big`, both on the encoder and decoder, indicated by areas with darker red around the diagonal. The prevalence of high similarity patterns among adjacent modules suggests a substantial degree of redundancy, and eliminating a module has a negligible impact on the final representations. On the other hand, we observe a distinct checkerboard pattern on the self-similarity matrices of the One Wide FFN model, where individual modules tend to exhibit lower similarity with their immediate neighbors than with their second neighbors (i.e., the neighbors of the neighbors). On the encoder, the checkerboard pattern emerges especially in the earlier modules while on the decoder, that pattern appears more consistently throughout the layers. This pattern gives an indication that our model is learning non-trivial

transformations of the input, leading to decreased redundancy within the network.

4.5 Other architectures and Languages

So far, all our experiments focused on the Transformer Big and on WMT22 EN →DE. In this section, we apply what we learned to other architectures and language pairs. We run experiments on the low resource language direction EN →RO and a large scale multilingual model.

For EN →DE, we apply our proposal to a Transformer Base model, a Deep Encoder Shallow Decoder model (Kasai et al., 2021), and a Decoder-Only model. For the Transformer Base, we observe an accuracy gain of 0.5 BLEU (2.2 BLEU over the vanilla `SharedEncNoDec` model) and an inference speedup of around 25%. In the Deep Encoder Shallow Decoder model, we observe a more modest accuracy gain of 0.2 BLEU points (0.9 BLEU over the vanilla `SharedEncNoDec` model). However, the inference speedup from dropping the decoder FFNs is minimal ($< 1\%$), which is expected because of the small depth of the decoder in this architecture.

Decoder-only models With the advent of Large Language Models (LLMs) like GPT (Brown et al., 2020), and PaLM (Chowdhery et al., 2022), a lot of effort has been put on decoder-only Transformer models. We train a decoder-only model on WMT22 EN →DE, as shown on Table 6. Due to the absence of an encoder, we are limited to applying a wide FFN on the decoder side. As in the other setups, we get an accuracy gain of +0.3 BLEU over the baseline decoder-only model (+1.7 BLEU over `ShareDec`), but the latency degrades by 12%. This is not surprising: due to the autoregressive nature of the decoder, increasing the size of its FFN has a bigger impact on speed.

Low-resource languages In EN →RO the accuracy of the One Wide FFN Model is only on par compared to the base model, even though it is a higher than the vanilla `SharedEncNoDec` model. We hypothesize that due to the low resource condition, our proposed model already reaches saturation as there are not that many salient textual patterns to be learned by the FFN.

Multilingual Finally, we observe the similar trend on the multilingual setup, where the One Wide FFN Model is +1.2 SPBLEU points more accurate than the baseline Transformer Big and +2.5 SPBLEU points more accurate than the vanilla `SharedEncNoDec`, this gain is significant in 79 out of 90 directions and when all tests sets are concatenated. Additionally, this large accuracy gain also comes with around 18% inference speed-up, consistent with our previous results.

5 Related Work

Weight pruning and parameter sharing are well-known techniques to reduce a model's footprint. Given the scale of the latest models (Chowdhery et al., 2022), there have been multiple efforts to prune neurons based on different automatic methods (Dalvi et al., 2020; Michel et al., 2019; Voita et al., 2019), sharing parameters efficiently (Ge et al., 2022; Reid et al., 2021), and factorizing certain components (Lan et al., 2020; Hu et al., 2022).

Neuron pruning methods often focus on finding and pruning redundant neurons through correlation methods (Dalvi et al., 2020), but also on how Transformer components like the multi-head attention can be pruned significantly due to model redundancy in the encoder or decoder either by checking the gradients salience (Michel et al., 2019) or a differentiable relaxation of the l_0 regularization at training time (Voita et al., 2019).

For parameter sharing, the Universal Transformer (Dehghani et al., 2019) proposed a model where all layers are shared (i.e., in effect it reduced the model to a single shared layer). Takase and Kiyono (2023) proposes finding an optimal configuration of shared layers in the encoder or decoder through different methods of sharing (in sequence, in cycle, or in reversed cycle) always keeping a specified number of final layers[9]. Similarly, Reid et al. (2021) proposes an approach where just the middle layers are shared, while the bottom and top layers are independent, and using a lower dimensionality for the embedding layer. Analogously, Ge et al. (2022) focus on minimizing the number of parameters and the number of calls to each parameters' group in order to optimise on-device models. They achieve this by sharing the encoder and decoder in a similar way to both previous methods, particularly by sharing all layer parameters in cycle like Takase and Kiyono (2023).

Previous works also focus on reducing the dimensionality of certain parameters, mostly through low rank factorization. Lan et al. (2020) decomposes the embedding layer into a lower rank em-

[9]See Appendix A for a detailed description and comparison.

	BLEU	CHRF	COMET	Speed	$\lvert\theta\rvert$ (%)
Transformer Base EN →DE	34.2	61.6	54.1	$116.3^{\pm 0.9}$	70M (100)
+ SharedEncNoDec FFN $d_{\text{ff}'} = 2,048$	$32.5^{\dagger}$	$60.1^{\dagger}$	50.0	$146.0^{\pm 1.6}$	47M (67)
+ SharedEncNoDec FFN $d_{\text{ff}'} = 24,576$	**34.7**	**61.8**	**55.6**	$146.8^{\pm 1.3}$	70M (100)
Transformer Decoder-Only EN →DE	35.8	62.8	57.7	$79.8^{\pm 1.9}$	202M (100)
+ ShareDec FFN $d_{\text{ff}'} = 4,096$	$34.4^{\dagger}$	$61.7^{\dagger}$	54.1	$79.7^{\pm 1.3}$	110M (48)
+ ShareDec FFN $d_{\text{ff}'} = 49,152$	**36.1**	**62.9**	**59.4**	$69.3^{\pm 0.2}$	202M (100)
Transformer Deep Enc. Shallow Dec. EN →DE	35.5	**62.4**	58.0	$230.1^{\pm 0.8}$	236M (100)
+ ShareEncNoDec FFN $d_{\text{ff}'} = 4,096$	$34.8^{\dagger}$	$61.6^{\dagger}$	55.4	$235.0^{\pm 0.5}$	127M (54)
+ ShareEncNoDec FFN $d_{\text{ff}'} = 57,344$	**35.7**	**62.4**	**58.9**	$233.5^{\pm 0.7}$	236M (100)
Transformer Base EN →RO	**22.9**	**52.9**	**50.9**	$119.3^{\pm 1.1}$	64M (100)
+ SharedEncNoDec FFN $d_{\text{ff}'} = 2,048$	$22.2^{\dagger}$	$52.5^{\dagger}$	45.8	$152.8^{\pm 1.4}$	41M (64)
+ SharedEncNoDec FFN $d_{\text{ff}'} = 24,576$	**22.9**	52.8	46.7	$150.6^{\pm 0.5}$	64M (100)
Transformer Big Multilingual	26.8	46.3	47.7	$94.6^{\pm 1.6}$	422M (100)
+ SharedEncNoDec FFN $d_{\text{ff}'} = 4,096$	$25.5^{\dagger}$	$45.1^{\dagger}$	40.8	$107.1^{\pm 1.4}$	330M (78)
+ SharedEncNoDec FFN $d_{\text{ff}'} = 49,152$	$\mathbf{28.0}^{\dagger}$	$\mathbf{47.3}^{\dagger}$	**50.7**	$111.5^{\pm 1.1}$	422M (100)

Table 6: Accuracy of One Wide FFN for EN →DE with Transformer Base, Decoder Only, and Deep Encoder Shallow Decoder on WMT22; for low resource EN →RO with Base version on WMT16, and multilingual with Transformer big on Flores. † implies the system is statistical significantly different at $p < 0.05$.

bedding matrix and a projection to the actual hidden size while also sharing all parameters across all layers. In addition to sharing parameters efficiently, Ge et al. (2022) proposes a lightweight decomposition of the FFN where instead of a single component there are 2 projections with a smaller dimensionality than vanilla Transformers. Our work is close to Ge et al. (2022) but instead of factorizing we explore sharing and full pruning of the FFN. In contrast with previous works, we also explore increasing the encoder FFN size while dropping the decoder's completely.

6 Conclusion

In this work, we studied the importance of the FFN in Transformer models. We analyzed the impact of removing and/or sharing the FFN across layers and found that, due to this component's redundancy, the model sizes can be substantially reduced with little impact on accuracy for Machine Translation. In particular, we found that sharing the FFN across all encoder layers while making it larger and removing it from the decoder layers leads to models that are more accurate and faster at inference.

Our findings are applicable across multiple settings, including decoder-only and multilingual models. In a low-resource setting the results are modest but our approach can still recover the baseline's performance with a faster inference.

Finally, we conducted a thorough similarity analysis between the vanilla Transformer and our proposed architectures, and found that the latter's internal representations do not differ significantly from the former's, except in that they are less redundant.

Limitations

In this work, our focus was Machine Translation. Although we expect the results to generalize to other sequence-to-sequence tasks, further experiments are needed, which we leave for future work.

Ethics Statement

One important consideration is the energy consumption for model training, which results in greenhouse emissions (Strubell et al., 2019). Our work uses existing datasets, and inherits some of the risks associated with them, such as privacy leakage (Carlini et al., 2021) and gender bias (Cho et al., 2019). Mitigation strategies such as those from Vanmassenhove et al. (2018) may be necessary.

Acknowledgements

We would like to thank Robin Schmidt, Matthias Sperber, and Stephan Peitz for their feedback and support in reviewing this work.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Yamini Bansal, Behrooz Ghorbani, Ankush Garg, Biao Zhang, Colin Cherry, Behnam Neyshabur, and Orhan Firat. 2022. Data scaling laws in NMT: The effect of noise and architecture. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 1466–1482. PMLR.

Angie Boggust, Brandon Carter, and Arvind Satyanarayan. 2022. Embedding Comparator: Visualizing Differences in Global Structure and Local Neighborhoods via Small Multiples. In *27th International Conference on Intelligent User Interfaces*, pages 746–766, Helsinki Finland. ACM.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.

Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 173–181, Florence, Italy. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Fahim Dalvi, Hassan Sajjad, Nadir Durrani, and Yonatan Belinkov. 2020. Analyzing redundancy in pretrained transformer models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4908–4926, Online. Association for Computational Linguistics.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Lukasz Kaiser. 2019. Universal transformers. In *International Conference on Learning Representations*.

Tao Ge, Si-Qing Chen, and Furu Wei. 2022. EdgeFormer: A parameter-efficient transformer for on-device seq2seq generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10786–10798, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Behrooz Ghorbani, Orhan Firat, Markus Freitag, Ankur Bapna, Maxim Krikun, Xavier Garcia, Ciprian Chelba, and Colin Cherry. 2022. Scaling laws for neural machine translation. In *International Conference on Learning Representations*.

Mitchell A Gordon, Kevin Duh, and Jared Kaplan. 2021. Data and parameter scaling laws for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5915–5922, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. 2005. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Cultural shift or linguistic drift? comparing two computational measures of semantic change.

In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121, Austin, Texas. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Jungo Kasai, Nikolaos Pappas, Hao Peng, James Cross, and Noah A. Smith. 2021. Deep encoder, shallow decoder: Reevaluating non-autoregressive machine translation. In *9th International Conference on Learning Representations, ICLR*, virtual. OpenReview.net.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Qian Lou, Ting Hua, Yen-Chang Hsu, Yilin Shen, and Hongxia Jin. 2022. Dictformer: Tiny transformer with shared dictionary. In *International Conference on Learning Representations*.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Telmo Pires, Robin Schmidt, Yi-Hsiu Liao, and Stephan Peitz. 2023. Learning language-specific layers for multilingual machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14767–14783, Toronto, Canada. Association for Computational Linguistics.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163, Valencia, Spain. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Machel Reid, Edison Marrese-Taylor, and Yutaka Matsuo. 2021. Subformer: Exploring weight sharing for parameter efficiency in generative transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4081–4090, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Robin Schmidt, Telmo Pires, Stephan Peitz, and Jonas Lööf. 2022. Non-autoregressive neural machine translation: A call for clarity. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2785–2799, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Sho Takase and Shun Kiyono. 2023. Lessons on parameter sharing across layers in transformers. In *Proceedings of The Fourth Workshop on Simple and*

Efficient Natural Language Processing (SustaiNLP), pages 78–90, Toronto, Canada (Hybrid). Association for Computational Linguistics.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy. Association for Computational Linguistics.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

A Custom Sharing of Multiple FFNs

There is a combinatorial number of ways of sharing $M < N$ FFNs within a module of N layers. Since this is prohibitive, we investigate the following strategies from Takase and Kiyono (2023):

- `Sequence`: assign one FFN for every M/N consecutive layers, forming a block pattern.

  $\text{FFN}_i(\cdot) \overset{\text{tied}}{=} \text{FFN}_{seq_m}(\cdot), \forall i : 1 \leq i \leq N, m = \lfloor (i-1)/(N/M) \rfloor$

- `Cycle`: stack M FFNs in an identical order, forming a repetitive checkerboard pattern.

  $\text{FFN}_i(\cdot) \overset{\text{tied}}{=} \text{FFN}_{cyc_m}(\cdot), \forall i : 1 \leq i \leq N, m = (i-1) \textbf{ modulo } M$

- `Cycle (Rev)`: stack M FFNs in a reverse order, forming a repetitive palindrome series.

  $\text{FFN}_i(\cdot) \overset{\text{tied}}{=} \text{FFN}_{cycrev_m}(\cdot), \forall i : 1 \leq i \leq N, m = N/M - i$

Note that we assume that N is an even number and divisible by N. `Cycle (Rev)` is only valid for $M = N/2$. The `EdgeFormer` (Ge et al., 2022) adopts `Cycle` with $M = 2$ for the encoder FFNs.

Table 7 shows the results of these strategies applied on the encoder. As references, we copy the results of the `Transformer Big` and `ShareEnc` from Table 2. Not only is the accuracy of `ShareEnc` similar to Takase and Kiyono (2023)'s strategies, but it also uses fewer parameters and is easier to extend.

Architecture	BLEU	$\lvert\theta\rvert$	(%)
`Transformer Big`	35.6	228M	(100)
+ `SharedEnc (M=1)`	35.4	186M	(82)
+ `Sequence M=2`	35.2	194M	(85)
+ `Sequence M=3`	35.3	202M	(88)
+ `Cycle M=2`	35.2	194M	(85)
+ `Cycle M=3`	35.5	202M	(88)
+ `Cycle Rev M=2`	35.2	194M	(85)
+ `Cycle Rev M=3`	35.5	202M	(88)

Table 7: Accuracy of different FFN sharing strategies on WMT22 EN → DE.

B Sharing or Dropping Attention

We report the results of sharing attention modules (either self, cross or both) across layers in Table 8. In contrast with the FFN, attention seems to play a more crucial role in the model's performance, as sharing the different attention mechanisms in both encoder and decoder causes a large accuracy drop across all settings, with the exception of sharing the decoder's cross attention and the encoder's self attention.

Encoder Self-Att	Decoder Self-Att	Decoder Cross-Att	BLEU	$\lvert\theta\rvert$	(%)
`Transformer Big`			35.6	228M	(100)
Shared	Shared	Shared	27.5	165M	(72)
Shared	Shared	Indiv.	27.6	186M	(82)
Shared	Indiv.	Indiv.	35.5	207M	(91)
Indiv.	Shared	Indiv.	26.5	207M	(91)
Indiv.	Shared	Shared	25.7	186M	(82)
Indiv.	Indiv.	Shared	35.5	207M	(91)

Table 8: BLEU scores on WMT 22 EN →DE when sharing the attention of both encoder and decoder (self and cross). Nomenclature follows Section 3 but with Self Attn an Cross Attn as the encoder/decoder's self attention and cross-attention (decoder), respectively.

C Details on Internal Representations Analysis

C.1 Raw Similarity Scores for Benchmarking

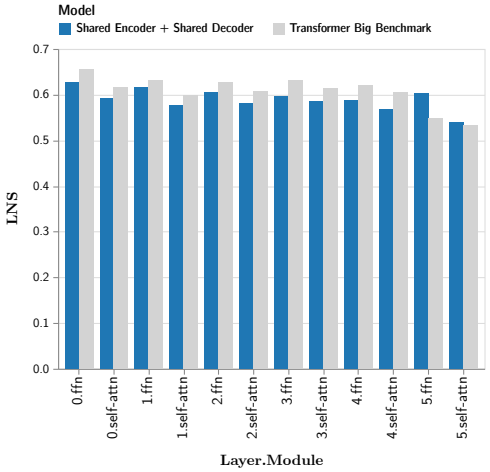
We establish a *benchmark* score for the expected similarity of our two metrics by comparing the baseline Transformer Big with identical models trained from different random seeds. Table 9 presents the raw similarity scores from which we compute the normalized scores presented in Table 5. As shown, the similarity between

Architecture	Encoder		Decoder	
	CKA	LNS	CKA	LNS
TransformerBig Seed 2	.96	.61	.94	.62
TransformerBig Seed 3	.96	.62	.95	.62
SharedEnc	.94	.58	.95	.62
SharedDec	.97	.62	.93	.59
SharedEncSharedDec	.95	.59	.94	.59
SharedEncDec	.94	.57	.93	.58
NoEnc	.87	.43	.95	.60
NoDec	.96	.60	.90	.54
ShareEncNoDec	.94	.59	.92	.55
ShareEncNoDec$^{d'_{ff}=41952}$	.94	.51	.89	.51

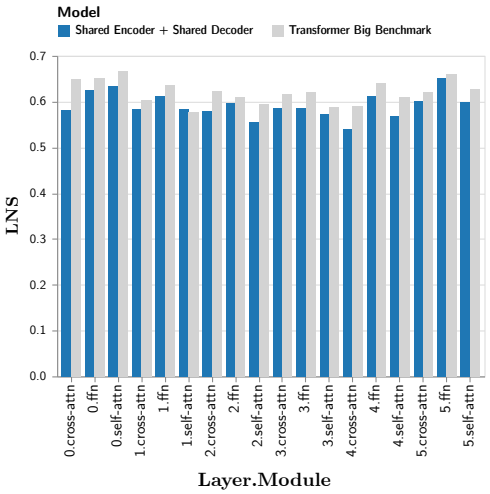
Table 9: Raw similarity of the representations of corresponding layer-modules of different architectures vs. the Transformer Big for WMT22 EN →DE. For *NoDec* configurations we compare the final output of the transformer layer as a whole as they have different submodules. The columns for shared and for dropped FFNs are highlighted in gray and blue respectively.

C.2 Layer-wise Analysis

In Table 5, we report the aggregated similarity scores across all layers of Transformer encoder and decoder. Here, we report a more fine-grained layer-wise similarity score mostly to showcase the reliability of the aggregated scores. In Figure 2, we plot layerwise LNS to study how similar the semantic information captured at each layer is to that of the baseline model at *every layer*. When LNS scores are high, the network is producing similar local neighborhoods for each sentence in our evaluation set. In particular, we are interested in comparing the benchmark LNS scores and those of SharedEncSharedDec at each layer. As shown, the layer-wise LNS scores of SharedEncSharedDec track the baseline scores at almost every layer, confirming the reliability of the aggregated score. We observe similar pattern for all the models that we evaluate in this paper.



(a) Encoder



(b) decoder

Figure 2: Layerwise LNS between SharedEncSharedDec and Transformer Big (blue bars). LNS between two versions of Transformer Big trained from different random initializations are shown by the grey bars to ground the comparison. FFN sharing does not dramatically change activations produced at each layer.

D Effect of batch size on decoding speed

In Section 4.3, we compared the decoding speeds of the One Wide FFN model and the Transformer Big, with a batch size of 1. In Table 10, we delve into how the decoding speed evolves as the batch size increases. As shown, the One Wide FFN model is faster for smaller batch sizes, but its advantage diminishes as the batch size increases, being slower than the Transformer Big for large batch sizes. We suspect this slowdown is due to the fact that the

Batch	Transformer Big	*One Wide FFN*	Speed-up (%)	*# batches*
1	$110.8^{\pm 1.2}$	$137.5^{\pm 1.1}$	24	2,047
2	$221.7^{\pm 14.3}$	$260.9^{\pm 6.5}$	18	1,024
4	$397.4^{\pm 8.0}$	$448.9^{\pm 2.0}$	13	512
8	$718.3^{\pm 8.0}$	$748.7^{\pm 10.6}$	4	256
16	$1,220.7^{\pm 56.2}$	$1,226.9^{\pm 17.2}$	1	128
32	$1,958.5^{\pm 112.4}$	$1,837.6^{\pm 15.3}$	−6	64
64	$1,319.1^{\pm 36.7}$	$1,259.0^{\pm 70.0}$	−5	32
128	$1,925.1^{\pm 64.8}$	$1,705.0^{\pm 62.3}$	−11	16
256	$2,312.1^{\pm 67.4}$	$1,976.5^{\pm 123.2}$	−15	8
512	$2,512.0^{\pm 50.1}$	$1,957.9^{\pm 32.6}$	−22	4

Table 10: Effect of batch size on decoding speed (in tokens/s) for the Transformer Big and *One Wide FFN* ($d_{\text{ff}'} = 49,152$). Δ is the percentage change in inference speed, and # batches is the number of batches used to evaluate. For large batch sizes, there are fewer batches (since the dataset size is fixed), which leads to higher variance in the measurements.

large FFN size requires higher peak memory, making the larger sizes non-optimal for this model.

A Benchmark for Evaluating Machine Translation Metrics on Dialects Without Standard Orthography

Noëmi Aepli[1] **Chantal Amrhein**[1,2] **Florian Schottmann**[2,3] **Rico Sennrich**[1,4]

[1]University of Zurich, [2]Textshuttle, [3]ETH Zurich, [4]University of Edinburgh

{naepli,sennrich}@cl.uzh.ch, {amrhein,schottmann}@textshuttle.com

Abstract

For sensible progress in natural language processing, it is important that we are aware of the limitations of the evaluation metrics we use. In this work, we evaluate how robust metrics are to non-standardized dialects, i.e. spelling differences in language varieties that do not have a standard orthography. To investigate this, we collect a dataset of human translations and human judgments for automatic machine translations from English to two Swiss German dialects. We further create a challenge set for dialect variation and benchmark existing metrics' performances. Our results show that existing metrics cannot reliably evaluate Swiss German text generation outputs, especially on segment level. We propose initial design adaptations that increase robustness in the face of non-standardized dialects, although there remains much room for further improvement. The dataset, code, and models are available here: https://github.com/textshuttle/dialect_eval

1 Introduction

As multilingual NLP models include more and more languages, the community's focus on low-resource languages has also grown. This not only includes languages for which we have "little data" but also language varieties and dialects which often pose additional challenges, especially if they do not have a standardized orthography. Recent work has shown some progress in classification tasks (e.g. Wang et al., 2021; Touileb and Barnes, 2021; Aepli and Sennrich, 2022) as well as generation tasks where such language varieties appear on the input side only (e.g. Zbib et al., 2012; Honnet et al., 2018; Alam et al., 2023). For these scenarios, we can use established evaluation schemes. However, for research towards NLP models *generating* language varieties, Sun et al. (2023) have shown that current evaluation metrics are not robust to translations into different dialects.

GSW	... ufere	Webs ii te	aa glueg e t	w ä rd e .
GSW	... ufere	Webs i te	ah gluegt	w e rd ä .
de	... auf einer	Webseite	angeschaut	werden.
en		... *viewed on a website.*		

Figure 1: Example sentence that shows the extent of spelling variability in language varieties, here Swiss German dialect (GSW), with German (de) and English (en) translations.

What their evaluation does not consider is that language varieties often lack a standardized orthography and do not adhere to consistent spelling rules. This implies that even *within* a single dialect, notable orthographic variations can be observed, as illustrated in the Swiss German example in Figure 1. The same utterance with a similar but different spelling would result in a high word error rate of $\frac{3}{4}$.

Many languages have multiple regional variants, such as Spanish (Mexican, Argentinean, etc.), French (Canadian, Belgian, etc.), or English (British, American, Australian, Indian, etc.), among others. Such language varieties exhibit various lexical, grammatical, and orthographical distinctions. Importantly, these differences are *standardized*, meaning that they adhere to specific spelling rules and conventions, albeit with variations specific to each variant. This suggests that if a neural metric is exposed to a sufficient amount of data encompassing various language varieties, it should be able to develop similar representations and provide comparable scores for a given sentence in different varieties. Sun et al. (2023) show that pre-training a metric on data from multiple dialects indeed makes metrics more inter-dialect robust.

However, for a substantial number of languages and language varieties, there exists no established standard orthography. Many regions exhibit a dialect continuum where language varieties lack precise boundaries, and each dialect displays a significant range of diversity within itself. Furthermore, when speakers write in their dialect, they follow

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 1045–1065
December 6–7, 2023.

their individual writing styles. Such kinds of variabilities, as can be observed in the example in Figure 1, are much less consistent and localized and will differ significantly between different writers. A metric designed to handle these kinds of varieties must be capable of addressing frequent spelling differences, which is considerably more challenging to learn solely from data compared to the standardized language variation differences mentioned in the previous paragraph.

In recent years, embedding-based metrics have gained increasing popularity (Sellam et al., 2020; Rei et al., 2020a) which – in theory – could be more appropriate for assessing non-standardized language varieties than string-based MT metrics like BLEU (Papineni et al., 2002) or chrF (Popović, 2015). However, these neural metrics are often not trained on the language varieties in question. Additionally, recent work showed that reference-based learned metrics still rely too much on subword overlap with the reference (Hanna and Bojar, 2021; Amrhein et al., 2022).

In this work, we follow Sun et al. (2023) and analyze the dialect robustness of machine translation metrics but specifically focus on non-standardized language varieties that were not seen during pretraining. Our contributions are:

- We collect a new dataset and design a challenge set for evaluating MT metrics on two Swiss German dialects.
- We benchmark existing string-based and neural metrics on our dataset and find that they are not reliable, especially on segment level.
- We propose initial adaptations to make metrics more robust for Swiss German but find that there is still a lot of room for improvement.

2 Related Work

There is a substantial amount of research on MT *into* language varieties (Scherrer, 2011b; Haddow et al., 2013; Fancellu et al., 2014; Hassani, 2017; Costa-jussà et al., 2018; Lakew et al., 2018; Myint Oo et al., 2019; Wan et al., 2020; Garcia and Firat, 2022). Most of these works exclusively evaluate with surface-level metrics like BLEU (Papineni et al., 2002) but some voice their concerns over a lack of reliable evaluation metrics (Kumar et al., 2021; Bapna et al., 2022).

Sun et al. (2023) confirm that existing machine translation evaluation metrics are not dialect-robust. They show that it is possible to train more robust metrics by including a language and dialect identification task in a second language model pre-training phase. While they focus on inter-dialect robustness between well-defined dialects, i.e. Brazilian and Iberian Portuguese, our study focuses on a setting where dialects lack standardized orthography. This absence of standardization introduces additional variability, resulting in distinct challenges and necessitating different solutions for MT systems, which need to generalize to often limited data; MT metrics, which need to be robust to spelling differences; and also meta-evaluation, which has its own challenges when collecting human assessments for dialects without standardized orthography as we outline in Section 3.1. To investigate how reliable MT metrics are for non-standardized varieties, we collect a new dataset with human translations and human judgments for MT outputs from English to two Swiss German dialects.

While other works also evaluate MT metrics on language varieties and dialects, Sun et al. (2023) is closest to our work: Alam et al. (2023) only look at language varieties on the source side and Riley et al. (2023) only evaluate language varieties for which a standard was included in the language model pretraining. Both studies also conclude that existing metrics are not robust to dialects. Riley et al. (2023) further propose a new automated lexical accuracy metric based on term dictionaries, similar to metrics used for automatic speech recognition (ASR) (Ali et al., 2017; Nigmatulina et al., 2020) which allow for more flexible string matching by using a look-up table of acceptable spellings. Riley et al.'s approach may work well if there is a limited set of term differences between dialects. However, such a metric is difficult to employ for language varieties without standardized spelling rules. Instead, we experiment with increasing dialect robustness by introducing character-level noise during metric training which has been shown to be useful for cross-lingual transfer to language varieties without standardized orthography (Aepli and Sennrich, 2022; Srivastava and Chiang, 2023; Blaschke et al., 2023).

3 Evaluation Data for Swiss German Dialects

While we focus on Swiss German because there are enough different MT systems that can be eval-

uated, Swiss German is by no means the only language where its varieties do not have standardized spelling. Many medium to high-resource languages like Arabic (Darwish et al., 2021) or Italian (Ramponi, 2022) include dialectal varieties that lack a standardized orthography. Additionally, this phenomenon extends to numerous low-resource settings (Bird, 2022), encompassing a wide array of language varieties across Africa (Adebara and Abdul-Mageed, 2022), Asia (Roark et al., 2020; Aji et al., 2022), Oceania (Solano et al., 2018) and the Americas (Littell et al., 2018; Mager et al., 2018). Historically, even many language varieties that now have a standardized orthography did not always have one, including English (Scragg, 1974). This makes our work on robust metrics for non-standardized dialects also relevant for NLP for historical texts.

To measure robustness against non-standardized dialects, we design two new datasets. With the first, we investigate how metrics behave in a realistic setup where we compare them against human judgments. The second is a challenge set that allows us to investigate score changes between different spellings and compare them to score changes when meaning is changed. This is inspired by similar experiments in Sun et al. (2023).

3.1 Human Judgement Data

In order to realistically evaluate machine translation metrics on Swiss German dialects, it is essential to obtain human-translated reference segments and human judgments for machine-translated translation hypotheses. Since no such data exists for Swiss German, we compile our dataset based on the English NTREX-128 data[1] (Federmann et al., 2022). We selected this dataset because it originates from a standard test set[2], already contains human translations into 128 languages including some regional variants, has a permissive license[3] and offers document context which is important for collecting reliable human judgments (Läubli et al., 2018; Toral et al., 2018).

Human reference translations: For the reference translations, we provided two Swiss German translators with the English NTREX-128 source data (i.e. 1997 sentences from 123 documents). Translators saw sentences in document context and were asked to translate them into their respective native dialects (i.e. Bern and Zurich region). We provided translators with simple instructions where we stated that they must not post-edit machine translation outputs to translate the texts.

Human judgment scores: The hypotheses come from ten machine translation systems translating from English to Bern dialect and ten systems translating from English to Zurich dialect. For each dialect, we include nine neural MT systems in our rating setup and one rule-based system.

The neural models are provided by Textshuttle. They are based on a standard Transformer architecture (Vaswani et al., 2017) trained using different amounts of data, making use of data augmentation techniques like backtranslation (Sennrich et al., 2016). Some of the systems use German as a pivot language. In collaboration with Textshuttle, we decided to evaluate models for which they expect noticeable translation differences and not to compare the nine models that they think would perform the best. The rule-based system works by morphosyntactically analyzing the standard German NTREX-128 translation of the English source and then sequentially applying a set of dialect-specific rewriting rules to generate Swiss German output. The system is described in detail in Scherrer (2011a). The system version used for this task operates word by word without taking syntax into account. Notably, this means that past tense and genitive forms produce unpredictable output because they would require larger changes in the sentence structure.

We translated the English NTREX-128 source data with each neural system and the German NTREX-128 translation with the rule-based systems and let native dialect speakers rate the outputs via Appraise[4] (Federmann, 2018), a framework for the evaluation of machine translation outputs. Raters only had access to the source for context because providing the reference could incentivize raters to "quickly compare the surface forms of translation against reference without understanding" (Freitag et al., 2022). Note that in order to mitigate dialect preference biases as documented by Riley et al. (2023) and Abu Farha and Magdy (2022), the translators and raters were all native speakers of the dialect they were asked to rate or translate into. We collected continuous Direct Assessment (DA) scores (Graham et al., 2013) where the slider

[1] https://github.com/MicrosoftTranslator/NTREX

[2] newstest2019 from the 2019 news translation shared task at WMT (Barrault et al., 2019)

[3] Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)

[4] https://github.com/AppraiseDev/Appraise

presented to the raters was annotated with Scalar Quality Metric (SQM) labels which increases the rating stability across annotators (Kocmi et al., 2022). Raters viewed segments in a document context and rated translations on the segment level as well as the document level. The document-level ratings are collected to enable future research on document-level metrics; in this study, we only focus on segment-level ratings.

Ideally, we would recruit professional translators for both the translation and the rating tasks. However, there exist no professional translators for Swiss German. Instead, we recruited translators and annotators from a pool of reliable candidates who already worked on similar Swiss German projects. To ensure the quality of the ratings we collect, we included control segments as implemented in Appraise. Based on this control, no raters needed to be excluded.

As Swiss German constitutes a dialect continuum, its various variations lack precise boundaries, and each dialect displays a significant range of diversity within itself. Consequently, during the recruitment process, we placed our trust in the annotators' self-identification of their native dialects. Furthermore, it is worth noting that all our contributors, comprising six women and five men, belong to younger generations, with raters ranging in age from 23 to 30, and translators aged 35 to 40, respectively. This age factor has an impact on their dialect. All translators and annotators were paid 30 CHF per hour for their work.

3.2 Challenge Set

As an additional evaluation, we compile a challenge set to directly pinpoint how robust metrics are to dialect variability. In the creation of this challenge set, we draw inspiration from the work of Sun et al. (2023), who propose measuring inter-dialect robustness by comparing metric scores between two language varieties and between one variety and a version with significant meaning changes. If segment pairs of the latter type are judged more or equally similar by a metric than those of the two varieties, Sun et al. (2023) argue the metric is not dialect-robust.

We build our challenge set from the collected data presented in the previous section. We filter for all MT hypotheses that humans rated as perfect (i.e. received a score of 100). If more than one unique hypothesis exists for a segment, we create all combinations of these hypotheses. For example, if four different machine translation outputs for the same source all receive a perfect human rating, this results in six pairs of semantically equivalent translation hypotheses that feature orthographic differences. For each pair, we then manually create a modified version of one of the hypotheses to change its meaning. Following Sun et al. (2023), we consider deletion, insertion, and substitution operations for introducing meaning changes which we randomly assign to each hypothesis pair. All changes are made either to a single word or if necessary a whole phrase. This process results in hypothesis triples as seen in this example:

A: S e chs Mitarbeiter s i wäg e Verletzige behandlet worde.
B: S ä chs Mitarbeiter s y wäg Verletzige behandlet worde.
Six members of staff have been treated for injuries.

C: Sechs Mitarbeiter si wäge Verletzige **beschtraft** worde.
*Six members of staff **were punished because of** injuries.*

Hypotheses A and B are semantically equivalent but exhibit spelling differences. Hypothesis C is very similar to hypothesis A on the surface level but differs significantly in meaning. During evaluation, metrics will have access to one of these hypotheses, as well as the reference and/or the source (depending on whether it is a reference-free or reference-based metric). We describe how we compare the different scores for these hypotheses in Section 4.3.

4 Experiment Setup

4.1 Benchmarking Existing Metrics

To document the performance of current MT metrics on dialects without a standard orthography, we evaluate the following metrics:

- **BLEU**[5] (Papineni et al., 2002), a string-based metric with a brevity penalty that calculates the word-level n-gram precision between a translation and one or multiple references.
- **chrF++**[6] (Popović, 2017), another string-based metric that provides a character n-gram, word unigram, and bigram F-score by computing overlaps between the hypothesis and reference translation.

[5]computed with SacreBLEU (Post, 2018), signature: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.0.
[6]computed with SacreBLEU (Post, 2018), signature: nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.3.0.

We expect surface-level, string-based metrics to perform badly on dialects without standard spelling rules as they are entirely based on overlap with a reference translation. These are also the metrics used by most works that explored text generation for language varieties without standardized orthography (e.g. Jeblee et al., 2014; Meftouh et al., 2015; Kumar et al., 2021). We further benchmark the following neural metrics:

- **COMET-20**[7] (Rei et al., 2020b) and **COMET-22**[8] (Rei et al., 2022), two reference-based neural metrics built on the COMET framework (Rei et al., 2020a). These are trained neural metrics that are built on top of a large, pre-trained language model and are fine-tuned on human judgment data from previous metric evaluation campaigns. COMET-20 is fine-tuned to predict DA scores. COMET-22 is an ensemble between a COMET-20-like model and a multi-task model that predicts segment-level Multidimensional Quality Metric (MQM) scores (Uszkoreit and Lommel, 2013) as well as word-level error tags.
- **COMET-20-QE**[9] (Rei et al., 2020b) and **COMET-Kiwi**[10] (Rei et al., 2022), two reference-free neural metrics for quality estimation. COMET-20-QE is trained similarly to COMET-20 and COMET-KIWI to COMET-22, but both versions do not have access to the reference during training on human judgments.

While these metrics go beyond surface-level comparisons to the reference due to their hidden representations and embedding-based nature, we expect that they still struggle to reliably evaluate translations into Swiss German for several reasons: First, no Swiss German data was included for pre-training the language model (XLM-R; Conneau et al., 2019) that is used as the basis for training COMET. Second, neural metrics are often fine-tuned on Standard German data which shares many similar words with Swiss German and could falsely bias metrics towards Standard German spelling. Third, reference-based metrics have been shown to still be influenced by surface overlap with the reference (Hanna and Bojar, 2021; Amrhein et al., 2022) which is a disadvantage in situations where numerous spelling variations exist.

[7] wmt20-comet-da
[8] wmt22-comet-da
[9] wmt20-comet-qe-da
[10] wmt22-cometkiwi-da

4.2 Developing Dialect-Robust Metrics

Similar to Sun et al. (2023), we also experiment with training more robust metrics but we focus on robustness against non-standardized dialects rather than inter-dialect robustness. The following list summarizes our metrics:

- **COMET-REF** and **COMET-QE**, a baseline trained as a reference to compare our modifications to because our COMET models differ slightly from COMET-20 and COMET-22 (see details below).
- **+gsw**, same as the baseline but the pre-trained model is fine-tuned on Swiss German data before the COMET models are fine-tuned on human judgment data. This is similar to the second pre-training phase for the inter-dialect-robust metric proposed in Sun et al. (2023). However, we do not include the additional language and dialect identification task during continued pre-training as we do not have dialect labels for the Swiss German pre-training data.
- **+noise**, same as the baseline but during the fine-tuning process on human judgment data we introduce character-level noise. This is inspired by previous work that showed that this method allows for better cross-lingual transfer to closely related languages (Aepli and Sennrich, 2022; Srivastava and Chiang, 2023). Blaschke et al. (2023) hypothesize that injecting noise into standard language data results in a similar tokenization rate as for unseen dialects. We apply noise injection to all languages within the COMET fine-tuning dataset that have an alphabetic writing system, therefore excluding languages like Chinese which were not considered in the original work introducing character-level noise. Following Aepli and Sennrich (2022), we inject character-level noise (essentially typos) into a random selection of 15% of the tokens within each sentence. Specifically, we alter, delete, or add one character per chosen token. We execute this process using the characters specific to the relevant language, taking into account all characters that occur more than 1,000 times in the respective dataset. We apply this noise

injection to all segments, including the source, translation, and reference segments.

We provide details of how we trained those models here:

Continued pre-training of XLM-R To expose our models to Swiss German data, we modify the encoder model upon which COMET models are usually based: XLM-RoBERTa[11] (Conneau et al., 2019). We continue the training of the XLM-R model on SwissCrawl[12] (Linder et al., 2020), a corpus containing 500K dialect sentences crawled from the web in late 2019. For the continued pre-training, we work with the Huggingface Transformers library[13] (Wolf et al., 2020), following the default configurations for language model fine-tuning which involves a training duration of three epochs.

Training COMET models We train COMET models using the official code base[14] with the default settings from version 2.0.2. We use the "regression model" configuration for the reference-based models and the "referenceless model" configuration for the reference-free models. Our models are trained on the direct assessment data collected by the organizers of the WMT news translation task spanning the years 2017 to 2021 (2021 as dev set)[15] (Bojar et al., 2017, 2018; Barrault et al., 2019, 2020; Akhbardeh et al., 2021). It is important to highlight that our models are not directly comparable to the original WMT shared task COMET models, for which the 2020 models were exclusively trained on data from 2017-2019 and the 2022 models used a different configuration.

4.3 Evaluation

We evaluate our metrics in five different ways. For the human judgment data, we compute two scores on system (sys) and two on segment (seg) level using the reference implementation from the WMT metrics shared task[16] (Freitag et al., 2022), except for *success rate* where we use our own implementation.

[11] xlm-roberta-base

[12] swisscrawl

[13] https://github.com/huggingface/transformers

[14] https://github.com/Unbabel/COMET

[15] https://github.com/Unbabel/COMET/tree/master/data

[16] https://github.com/google-research/mt-metrics-eval

System level The *pairwise accuracy* as defined by Kocmi et al. (2021), measures the accuracy with which a metric agrees with human preference between pairs of systems where the human ratings are significantly different according to a two-sided Wilcoxon test. Note that the score difference between the two systems is not important in this analysis. Furthermore, we provide results for the *sys-level Pearson correlation*, quantifying the strength of the linear relationship between metrics and human judgment scores for systems.

Segment level At the segment level, our evaluation includes the *seg-level accuracy* with an optimized tie threshold, which resembles a global accuracy but also acknowledges metrics for correctly predicting tied human judgment scores (Deutsch et al., 2023). Further, we present the *seg-level Kendall correlation*, akin to pairwise accuracy but employing a distinct normalization technique.

Challenge set For the challenge set, we compute the *success rate* (seg level) following Sun et al. (2023). This measures the accuracy with which a metric assigns more similar scores (s) to two equivalent translations A and B compared to a version with a semantic change C. Consequently, a metric is considered robust to non-standardized dialects for a segment if the score difference between s_A and s_B is smaller than the score difference between s_C and either s_A or s_B (depending on which score is smaller):

$$|s_A - s_B| < min(s_A, s_B) - s_C \quad (1)$$

5 Results

Table 1 provides a comprehensive summary of our results with scores for existing metrics (top), COMET models trained for this work (bottom), system-level evaluations (left), and segment-level evaluations (right). Additional results can be found in the appendices. Appendix A.1 contains results related to the incorporation of additional languages in the pre-training process, Appendix B presents an evaluation of performance on an official WMT benchmark, and Appendix C presents pairwise accuracy plots for our metrics.

Existing vs. GSW metrics As expected, the surface-level metrics perform worse than trained metrics in almost all evaluations. Our baseline metrics often perform a bit worse than the existing COMET metrics, this is particularly true for

	system-level			segment-level					
	pairwise accuracy	**Pearson correlation**		**tie-optim. accuracy**		**Kendall correlation**		**success rate**	
		BE	ZH	BE	ZH	BE	ZH	BE	ZH
BLEU	0.740	0.728	0.587	0.544	0.560	0.142	0.163	0.135	0.194
chrF	0.753	0.806	0.665	0.486	0.478	0.076	0.079	0.121	0.145
COMET-20	0.766	0.849	0.816	0.565	0.583	0.205	0.227	0.250	0.298
COMET-22	0.766	0.897	0.901	0.570	0.587	0.184	0.212	0.243	0.306
COMET-20-QE	0.675	0.875	0.872	0.508	0.516	0.134	0.134	0.131	0.161
COMET-KIWI	0.636	0.952	0.876	0.536	0.533	0.146	0.142	0.240	0.290
COMET-REF	0.740	0.864	0.793	0.567	0.570	0.180	0.194	0.221	0.234
+ gsw	0.792	**0.906**	0.862	0.611	0.627	**0.286**	**0.317**	0.320	0.347
+ noise	0.727	**0.940**	**0.903**	0.561	0.567	**0.223**	**0.233**	0.237	0.290
+ gsw + noise	0.792	**0.917**	**0.868**	0.597	0.621	**0.271**	**0.304**	0.287	0.323
COMET-QE-KIWI	0.636	0.781	0.689	0.486	0.507	0.104	0.099	0.127	0.145
+ gsw	0.844	**0.978**	**0.987**	0.595	0.587	**0.257**	**0.283**	0.292	0.298
+ noise	0.675	**0.915**	**0.817**	0.524	0.528	**0.154**	**0.158**	0.149	0.177
+ gsw + noise	**0.896**	**0.968**	**0.981**	0.582	0.596	**0.246**	**0.269**	0.273	0.274

Table 1: Results for the baselines metrics (above) and our trained metrics (below) on system level (left) and segment level (right). Darker shades indicate lower scores. Bold denotes statistically significant improvement compared to their respective baselines COMET-REF or COMET-QE-KIWI. There is no information about significance for tie-optim. accuracy (columns 4-5) and success rate (columns 8-9). Note that BE and ZH represent the abbreviations for the two Swiss German (GSW) dialect regions under consideration.

our reference-free model. However, continued pre-training on Swiss German data improves their performance considerably and they strongly outperform existing metrics. This highlights the importance of the model to have seen the target language (variety) during the language model pre-training. It also shows that metrics can be extended to include new languages and language varieties with limited effort although this impacts their performance on other language pairs as we show in Appendix B. Continued pre-training on multiple languages and language varieties can mitigate this effect (see Appendix A.1).

Noise injection While continued LM pre-training on Swiss German data generally outperforms noise injection during task fine-tuning, we still see gains over the baselines. This suggests that metrics that were trained on noised data are more robust to unseen language (varieties) and may be a good strategy for language (varieties) without sufficient data for continued pre-training. Combining both continued pre-training and noise injection generally does not lead to further improvements.

Reference-based vs reference-free While both types of metrics perform similarly with continued pre-training on Swiss German, both existing reference-free metrics perform worse than the existing reference-based metrics in the segment-level evaluations. Since these metrics did not see any Swiss German during the pre-training phase, having access to the reference as an anchor might help the reference-based metrics for unseen languages. Amrhein et al. (2022) reported a similar finding where the reference acted as an anchor when metrics were used to identify copied source sentences.

Challenge set The success rate for all metrics is extremely low. Metrics assign more similar scores to a hypothesis with a semantic change than to a different translation hypothesis in the majority of cases. Again, continued pre-training on Swiss German results in the best metric performance. However, even these scores are lower than a random success rate of 50% by far. Our findings highlight that even though system-level correlations may seem convincing, none of the metrics studied in this work are robust to non-standardized dialect variations.

Since our results show that there is still significant room for improvement toward metric robustness to non-standardized language varieties, we provide suggestions for future work.

6 Open Questions

We hope that our benchmark inspires more work on robust evaluation metrics for language varieties in the future. In this section, we list several directions we think are worthwhile exploring:

Expanding the benchmark: We were not able to include additional language varieties in our benchmark at the time because we could not find enough *different* machine translation systems that translate into these varieties. While we recognize that without reliable metrics this is a "chicken-and-egg" problem, we still advocate for more MT research that focuses on translating *into* language varieties. Expanding our benchmark would not only allow us to draw more general conclusions but would also help with sample size for the *pairwise accuracy* analysis (Kocmi et al., 2021) since we find that a large number of systems are required for confident results.

More focus on segment level: Segment-level metric scores tend to be much less correlated with human judgments when contrasted with system-level correlations (Freitag et al., 2022) and have also been shown to be unreliable in downstream tasks (Moghe et al., 2023). We hope that future work aimed at enhancing metric performance on our challenge set will also contribute to greater metric reliability on segment level in general, as over-reliance on reference overlap is also a problem for languages with standardized spelling (Hanna and Bojar, 2021; Amrhein et al., 2022).

Training neural metrics that model character-level similarities: A segment in a dialect often resembles a reference in certain characters only rather than in full words (see Figure 1 as an example). As the underlying language models of neural evaluation metrics use a fixed tokenization scheme that was learned on text that likely does not include many examples of language varieties, these similarities might be hard to account for by the neural metric. Thus, we believe that character-based language models, such as Canine (Clark et al., 2022), could provide a better basis for neural evaluation metrics to model character-level similarities.

7 Conclusion

We evaluated the reliability of machine translation metrics when evaluating dialects without standard orthographies. As part of this work, we collected a new dataset consisting of human translations, human judgments, and a challenge set from English to two Swiss German dialects. We benchmark several existing metrics and find that they are not robust to variation featured by non-standardized dialects. Based on this finding, we explore several modifications that allow us to train metrics that are more robust towards spelling variation. Our results show that there is still a lot of room for improvement and we offer a set of recommendations for future work on dialect robust metrics.

Limitations

The goal of this work is to evaluate and develop machine translation metrics that take into account the spelling variability of dialects and languages without established writing norms. We recognize that evaluating metrics on varieties from different languages would help generalize our results. However, we were not able to find enough differing machine translation systems that translate *into* the same language variety for other languages. Therefore, we had to limit this study to two Swiss German dialects. We hope to include further language varieties in our benchmark in the future (when such machine translation systems become available) to encourage research toward metrics that are reliable for many non-standardized language varieties.

We did our best to avoid dialectal preference bias within our annotators by selecting only annotators who consider themselves native speakers of the respective dialect. However, as Swiss German is a dialect continuum, this can only be controlled to a certain degree.

Ethics Statement

This work includes the compilation of a new dataset as a test set for evaluating various machine translation metrics. All translators and annotators were compensated at a rate of 30 CHF per hour. Our dataset is based on a publicly available dataset and will be released under the same license for future use. **Intended use:** The dataset and the models resulting from this work are intended to be used by the research community to evaluate machine translation metrics.

Acknowledgements

We thank Yves Scherrer for providing the rule based systems and helpful comments. Furthermore, we thank Annette Rios, Tom Kocmi, Mathias Müller, and the anonymous reviewers for their valuable inputs. We are also grateful to the Swiss German raters and translators for their important contribution. This work was supported by the Swiss National Science Foundation (project nos. 191934 & 176727), Textshuttle, and the Department of Computational Linguistics at the University of Zurich.

References

Ibrahim Abu Farha and Walid Magdy. 2022. The effect of Arabic dialect familiarity on data annotation. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 399–408, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards afrocentric NLP for African languages: Where we are and where we can go. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3814–3841, Dublin, Ireland. Association for Computational Linguistics.

Noëmi Aepli and Rico Sennrich. 2022. Improving zero-shot cross-lingual transfer between closely related languages by injecting character-level noise. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4074–4083, Dublin, Ireland. Association for Computational Linguistics.

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Md Mahfuz Ibn Alam, Sina Ahmadi, and Antonios Anastasopoulos. 2023. Codet: A benchmark for contrastive dialectal evaluation of machine translation. *arXiv preprint arXiv:2305.17267*.

Ahmed Ali, Preslav Nakov, Peter Bell, and Steve Renals. 2017. Werd: Using social text spelling variants for evaluating dialectal speech recognition. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 141–148.

Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2022. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Nikhil Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Saldinger Axelrod, Jason Riesa, Yuan Cao, Mia Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apu Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Richard Hughes. 2022. Building machine translation systems for the next thousand languages. Technical report, Google Research.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Steven Bird. 2022. Local languages, third spaces, and other high-resource scenarios. In *Proceedings of the*

60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7817–7829, Dublin, Ireland. Association for Computational Linguistics.

Verena Blaschke, Hinrich Schütze, and Barbara Plank. 2023. Does manipulating tokenization aid cross-lingual transfer? a study on POS tagging for non-standardized languages. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 40–54, Dubrovnik, Croatia. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Marta R. Costa-jussà, Marcos Zampieri, and Santanu Pal. 2018. A neural approach to language variety translation. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 275–282, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Kareem Darwish, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T. Al-Natsheh, Houda Bouamor, Karim Bouzoubaa, Violetta Cavalli-Sforza, Samhaa R. El-Beltagy, Wassim El-Hajj, Mustafa Jarrar, and Hamdy Mubarak. 2021. A panoramic survey of natural language processing in the arab world. *Commun. ACM*, 64(4):72–81.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Modifying kendall's tau for modern metric meta-evaluation. *arXiv preprint arXiv:2305.14324*.

Federico Fancellu, Andy Way, and Morgan O'Brien. 2014. Standard language variety conversion for content localisation via SMT. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pages 143–149, Dubrovnik, Croatia. European Association for Machine Translation.

Christian Federmann. 2018. Appraise evaluation framework for machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Xavier Garcia and Orhan Firat. 2022. Using natural language prompts for machine translation. *arXiv preprint arXiv:2202.11822*.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Barry Haddow, Adolfo Hernández, Friedrich Neubarth, and Harald Trost. 2013. Corpus development for machine translation between standard and dialectal varieties. In *Proceedings of the Workshop on Adaptation of Language Resources and Tools for Closely Related Languages and Language Variants*, pages 7–14, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Michael Hanna and Ondřej Bojar. 2021. A fine-grained analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online. Association for Computational Linguistics.

Hossein Hassani. 2017. Kurdish interdialect machine translation. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 63–72, Valencia, Spain. Association for Computational Linguistics.

Pierre-Edouard Honnet, Andrei Popescu-Belis, Claudiu Musat, and Michael Baeriswyl. 2018. Machine translation of low-resource spoken dialects: Strategies for normalizing Swiss German. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Serena Jeblee, Weston Feely, Houda Bouamor, Alon Lavie, Nizar Habash, and Kemal Oflazer. 2014. Domain and dialect adaptation for machine translation into Egyptian Arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 196–206, Doha, Qatar. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Sachin Kumar, Antonios Anastasopoulos, Shuly Wintner, and Yulia Tsvetkov. 2021. Machine translation into low-resource language varieties. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 110–121, Online. Association for Computational Linguistics.

Surafel Melaku Lakew, Aliia Erofeeva, and Marcello Federico. 2018. Neural machine translation into language varieties. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 156–164, Brussels, Belgium. Association for Computational Linguistics.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Lucy Linder, Michael Jungo, Jean Hennebert, Claudiu Cristian Musat, and Andreas Fischer. 2020. Automatic creation of text corpora for low-resource languages from the internet: The case of swiss german. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2706–2711, Marseille, France. European Language Resources Association.

Patrick Littell, Anna Kazantseva, Roland Kuhn, Aidan Pine, Antti Arppe, Christopher Cox, and Marie-Odile Junker. 2018. Indigenous language technologies in Canada: Assessment, challenges, and successes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2620–2632, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Karima Meftouh, Salima Harrat, Salma Jamoussi, Mourad Abbas, and Kamel Smaili. 2015. Machine translation experiments on PADIC: A parallel Arabic DIalect corpus. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 26–34, Shanghai, China.

Nikita Moghe, Tom Sherborne, Mark Steedman, and Alexandra Birch. 2023. Extrinsic evaluation of machine translation metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13060–13078, Toronto, Canada. Association for Computational Linguistics.

Thazin Myint Oo, Ye Kyaw Thu, and Khin Mar Soe. 2019. Neural machine translation between Myanmar (Burmese) and Rakhine (Arakanese). In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 80–88, Ann Arbor, Michigan. Association for Computational Linguistics.

Iuliia Nigmatulina, Tannon Kew, and Tanja Samardzic. 2020. ASR for non-standardised languages with dialectal variation: the case of Swiss German. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 15–24, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*,

pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Alan Ramponi. 2022. Nlp for language varieties of italy: Challenges and the path forward. *arXiv preprint arXiv:2209.09757*.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020a. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020b. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. Association for Computational Linguistics.

Parker Riley, Timothy Dozat, Jan A. Botha, Xavier Garcia, Dan Garrette, Jason Riesa, Orhan Firat, and Noah Constant. 2023. FRMT: A benchmark for few-shot region-aware machine translation. *Transactions of the Association for Computational Linguistics*, 11:671–685.

Brian Roark, Lawrence Wolf-Sonkin, Christo Kirov, Sabrina J. Mielke, Cibu Johny, Isin Demirsahin, and Keith Hall. 2020. Processing South Asian languages written in the Latin script: the Dakshina dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2413–2423, Marseille, France. European Language Resources Association.

Yves Scherrer. 2011a. Morphology generation for swiss german dialects. In *Systems and Frameworks for Computational Morphology*, pages 130–140, Berlin, Heidelberg. Springer Berlin Heidelberg.

Yves Scherrer. 2011b. Syntactic transformations for Swiss German dialects. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 30–38, Edinburgh, Scotland. Association for Computational Linguistics.

D. Scragg. 1974. *A History of English Spelling*. Manchester University Press, United Kingdom.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rolando Coto Solano, Sally Akevai Nicholas, and Samantha Wray. 2018. Development of natural language processing tools for Cook Islands Māori. In *Proceedings of the Australasian Language Technology Association Workshop 2018*, pages 26–33, Dunedin, New Zealand.

Aarohi Srivastava and David Chiang. 2023. Fine-tuning BERT with character-level noise for zero-shot transfer to dialects and closely-related languages. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 152–162, Dubrovnik, Croatia. Association for Computational Linguistics.

Jiao Sun, Thibault Sellam, Elizabeth Clark, Tu Vu, Timothy Dozat, Dan Garrette, Aditya Siddhant, Jacob Eisenstein, and Sebastian Gehrmann. 2023. Dialect-robust evaluation of generated text. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6010–6028, Toronto, Canada. Association for Computational Linguistics.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Samia Touileb and Jeremy Barnes. 2021. The interplay between language similarity and script on a novel multi-layer Algerian dialect corpus. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3700–3712, Online. Association for Computational Linguistics.

Hans Uszkoreit and Arle Lommel. 2013. Multidimensional quality metrics: A new unified paradigm for human and machine translation quality assessment. *Localization World, London*, pages 12–14.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yu Wan, Baosong Yang, Derek F. Wong, Lidia S. Chao, Haihua Du, and Ben C.H. Ao. 2020. Unsupervised neural dialect translation with commonality and diversity modeling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9130–9137.

Xinyi Wang, Yulia Tsvetkov, Sebastian Ruder, and Graham Neubig. 2021. Efficient test time adapter ensembling for low-resource language varieties. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 730–737, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine translation of Arabic dialects. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 49–59, Montréal, Canada. Association for Computational Linguistics.

A Appendix

A.1 Mixed Continued Pre-training

In our main experiments in Section 5, we evaluate continued language model pre-training only on Swiss German data. While this increases the performance on our benchmark, it remains unclear whether this leads to a "specialized" metric that does not perform well on other language pairs. We will evaluate this in the next section, but first, we introduce a set of contrastive models that are less specialized to Swiss German. Continued pre-training for contrastive models involves incorporating mixed data from five languages apart from Swiss German, namely: German (de), English (en), French (fr), Hindi (hi), and Chinese (zh). We train one metric based on XLM-R with continued pre-training only on these five languages ("5 langs"), and another one where we also add GSW to the training data ("6 langs"). For both settings, we also test character-level noise in the COMET fine-tuning step, as described in Section 4.2. The data for the five additional languages is sourced from the CC-100 corpus[17] (Wenzek et al., 2020), which is a reconstructed version of XLM-R's training dataset. Specifically, we utilize the first 100,000 sentences from the training data of each language.

Table 2 shows the results we obtained from incorporating mixed data into the continued LM pre-training. We see a similar effect as when continuing the pre-training only on GSW in the main results in Section 5. The performance of the metrics increases in all evaluations. Comparing these results to the metric where we only continued pre-training on Swiss German (+6 langs vs. +gsw), the results are comparable and often not significantly different. In the next section, we investigate how these metrics behave on other language pairs.

B Correlations on WMT Benchmarks

As discussed in the previous section, we evaluate the performance of our metrics on an official WMT benchmark to monitor their performance on language pairs that do not involve Swiss German. To do this, we reproduce the evaluations from the WMT 2022 metrics task (Freitag et al., 2022) for a subset of language pairs. We evaluate on the following five language pairs:

[17] https://data.statmt.org/cc-100/

	system-level			segment-level					
	pairwise accuracy	**Pearson correlation**		**tie-optim. accuracy**		**Kendall correlation**		**success rate**	
		BE	ZH	BE	ZH	BE	ZH	BE	ZH
COMET-REF	0.740	0.864	0.793	0.567	0.570	0.180	0.194	0.221	0.234
+ noise	0.727	**0.940**	**0.903**	0.561	0.567	**0.223**	**0.233**	0.237	0.290
+ gsw	0.792	**0.906**	0.862	0.611	0.627	**0.286**	**0.317**	0.320	0.347
+ gsw + noise	0.792	**0.917**	**0.868**	0.597	0.621	**0.271**	**0.304**	0.287	0.323
+ 5 langs	0.766	0.877	0.774	0.561	0.583	**0.212**	**0.230**	0.235	0.274
+ 5 langs + noise	0.766	**0.938**	**0.890**	0.570	0.593	**0.241**	**0.256**	0.265	0.290
+ 6 langs	0.805	**0.932**	**0.887**	0.592	0.616	**0.286**	**0.316**	0.357	0.452
+ 6 langs + noise	0.779	**0.956**	**0.917**	0.599	0.622	**0.282**	**0.311**	0.323	0.379
COMET-QE-KIWI	0.636	0.781	0.689	0.486	0.507	0.104	0.099	0.127	0.145
+ noise	0.675	**0.915**	**0.817**	0.524	0.528	**0.154**	**0.158**	0.149	0.177
+ gsw	0.844	**0.978**	**0.987**	0.595	0.587	**0.257**	**0.283**	0.292	0.298
+ gsw + noise	**0.896**	**0.968**	**0.981**	0.582	0.596	**0.246**	**0.269**	0.273	0.274
+ 5 langs	0.610	0.758	0.773	0.514	0.505	**0.134**	**0.135**	0.164	0.202
+ 5 langs + noise	0.701	0.898	**0.831**	0.513	0.521	**0.178**	**0.184**	0.166	0.266
+ 6 langs	0.831	**0.985**	**0.984**	0.583	0.605	**0.261**	**0.284**	0.304	0.331
+ 6 langs + noise	0.870	**0.983**	**0.983**	0.579	0.591	**0.251**	**0.269**	0.284	0.323

Table 2: Results for systems with continued pre-training only on Swiss German (+ gsw), on 5 other languages (+ 5 langs) and the same languages including Swiss German (+ 6 langs). Darker shades indicate lower scores. Bold denotes statistically significant improvement compared to their respective baselines COMET-REF or COMET-QE-KIWI. There is no information about significance for tie-optim. accuracy (columns 4-5) and success rate (columns 8-9). Note that BE and ZH represent the abbreviations for the two Swiss German (GSW) dialect regions under consideration.

- **en-de:** evaluation against MQM ratings collected specifically for the metrics shared task.
- **en-zh:** evaluation against MQM ratings collected specifically for the metrics shared task.
- **de-en:** evaluation against reference-based DA scores collected for the translation shared task.
- **cs-uk:** evaluation against DA + SQM scores collected for the translation shared task.
- **en-liv:** evaluation against DA + SQM scores collected for the translation shared task.

Note that all these languages except for Livonian (liv) are part of the CC-100 corpus[18] (Wenzek et al., 2020). Consequently, they form a part of the training dataset for XLM-R and are thus included in the COMET models. Moreover, English (en), German (de), and Chinese (zh) were incorporated into the mixed continued pre-training, as explained in Section A.1. Lastly, all the languages mentioned above, with the exception of Ukrainian (uk) and Livonian (liv; a language of Latvia), are included in the COMET training data.

This evaluation allows us to assess the effects of our modifications both on language pairs that were included during COMET training, during continued LM pre-training, and those that were not.

The results are shown in the following Tables: 3 (system-level Pearson correlation), 4 (segment-level accuracy), and 5 (segment-level Kendall). We do not report pairwise accuracy here because they cannot be directly compared with the WMT22 results, given that we have only included a subset of the language pairs. Versions of COMET-ref that were continued pretrained on Swiss German data demonstrate comparable or improved performance compared to the baseline metrics. In contrast, continued pretrained COMET-qe performs worse. When examining individual languages, we observe that fine-tuning is advantageous for translations into Livonian (liv), which is the only language in our selection not included in XLMR. Conversely, for translations into English, continued pretrained systems, particularly COMET-qe, tend to perform slightly worse.

C Pairwise Comparison Plots

In the subsequent plots displayed in Figures 2 (existing metrics), 3 (our trained COMET-ref metrics), and 4 (our trained COMET-qe metrics), every point represents a difference in average human judgment (y-axis) and a difference in automatic metric (x-axis) over a pair of systems. Metrics disagree with human ranking for system pairs in pink quadrants. These plots follow the example of Figure 1 in (Kocmi et al., 2021).

[18] https://data.statmt.org/cc-100/

sys-level Pearson correlation	de-en	en-de	en-zh	en-liv	cs-uk
BLEU	0.353	0.178	0.065	-0.575	0.890
chrF++	0.356	0.304	0.203	-0.517	0.925
COMET-20	0.424	0.876	0.744	0.893	0.985
COMET-22	0.450	0.873	0.756	-0.517	0.989
COMET-20-QE	0.443	0.577	0.752	0.564	0.953
COMET-KIWI	0.421	0.748	0.767	-0.563	0.987
COMET-ref	0.423	0.888	0.626	0.909	0.992
+ noise	0.420	**0.931**	0.618	0.912	0.991
+ gsw	0.410	0.904	0.450	0.693	0.983
+ gsw + noise	0.407	0.930	0.375	0.610	0.964
+ 5 langs	0.412	0.897	0.656	0.826	0.993
+ 5 langs + noise	0.415	**0.933**	0.658	0.689	0.991
+ 6 langs	0.417	0.908	0.636	0.892	0.992
+ 6 langs + noise	0.413	**0.951**	0.626	0.627	0.989
COMET-qe	0.384	0.453	0.639	0.598	0.954
+ noise	0.398	0.464	0.659	0.589	0.961
+ gsw	0.365	0.300	0.444	0.806	0.874
+ gsw + noise	0.387	0.354	0.446	0.859	0.893
+ 5 langs	0.371	0.434	0.650	0.621	0.923
+ 5 langs + noise	0.377	0.429	0.667	0.639	0.939
+ 6 langs	0.372	0.424	0.657	0.694	0.921
+ 6 langs + noise	0.380	0.440	0.640	0.725	0.939

Table 3: System-level Pearson correlation scores for baseline metrics (above) and our trained metrics (below) on a subset of language pairs from the WMT 2022 metrics task. Bold denotes statistically significant improvement compared to their respective baselines COMET-REF or COMET-QE-KIWI, underlined denotes statistically significant decline.

seg-level tie-optim. **accuracy**	de-en	en-de	en-zh	en-liv	cs-uk
BLEU	0.394	0.539	0.096	0.319	0.490
chrF++	0.391	0.545	0.352	0.237	0.466
COMET-20	0.439	0.580	0.466	0.589	0.563
COMET-22	0.437	0.584	0.468	0.368	0.567
COMET-20-QE	0.442	0.566	0.460	0.513	0.556
COMET-KIWI	0.412	0.580	0.470	0.338	0.567
COMET-ref	0.439	0.565	0.462	0.540	0.556
+ noise	0.434	0.556	0.458	0.615	0.564
+ gsw	0.434	0.551	0.470	0.507	0.542
+ gsw + noise	0.432	0.543	0.470	0.453	0.531
+ 5 langs	0.444	0.570	0.471	0.593	0.543
+ 5 langs + noise	0.428	0.553	0.478	0.500	0.552
+ 6 langs	0.445	0.567	0.475	0.461	0.551
+ 6 langs + noise	0.430	0.560	0.483	0.523	0.545
COMET-qe	0.433	0.550	0.470	0.545	0.555
+ noise	0.436	0.561	0.470	0.520	0.544
+ gsw	0.445	0.546	0.472	0.583	0.518
+ gsw + noise	0.441	0.552	0.463	0.505	0.500
+ 5 langs	0.445	0.561	0.467	0.522	0.530
+ 5 langs + noise	0.439	0.550	0.462	0.526	0.520
+ 6 langs	0.443	0.555	0.480	0.520	0.528
+ 6 langs + noise	0.453	0.552	0.470	0.517	0.526

Table 4: Segment-level accuracy scores (the darker the lower) for baseline metrics (above) and our trained metrics (below) on a subset of language pairs from the WMT 2022 metrics task. There is no information about significance.

seg-level Kendall correlation	de-en	en-de	en-zh	en-liv	cs-uk
BLEU	0.009	0.169	0.032	-0.158	0.133
chrF++	0.007	0.146	0.056	-0.158	0.086
COMET-20	0.018	0.319	0.141	0.208	0.280
COMET-22	0.019	0.343	0.137	-0.111	0.295
COMET-20-QE	0.022	0.234	0.123	0.126	0.254
COMET-KIWI	0.016	0.231	0.123	-0.147	0.281
COMET-ref	0.015	0.320	0.139	0.213	0.267
+ noise	0.019	0.310	0.125	0.165	0.251
+ gsw	0.016	0.293	0.120	0.096	0.225
+ gsw + noise	0.020	0.298	0.101	0.059	0.213
+ 5 langs	0.017	0.316	0.131	0.127	0.252
+ 5 langs + noise	0.018	0.321	0.128	0.095	0.238
+ 6 langs	0.017	0.309	0.133	0.140	0.246
+ 6 langs + noise	**0.018**	0.309	0.126	0.070	0.234
COMET-qe	0.017	0.225	0.121	0.152	0.235
+ noise	0.014	**0.228**	0.114	0.137	0.217
+ gsw	0.013	0.178	0.093	0.146	0.161
+ gsw + noise	0.013	0.182	0.094	0.102	0.162
+ 5 langs	0.020	0.214	0.115	0.145	0.214
+ 5 langs + noise	0.019	0.217	0.117	0.142	0.203
+ 6 langs	0.015	0.216	0.117	0.167	0.198
+ 6 langs + noise	0.016	0.212	0.114	0.147	0.186

Table 5: Segment-level Kendall correlation scores for baseline metrics (above) and our trained metrics (below) on a subset of language pairs from the WMT 2022 metrics task. Bold denotes statistically significant improvement compared to their respective baselines COMET-REF or COMET-QE-KIWI, underlined denotes statistically significant decline.

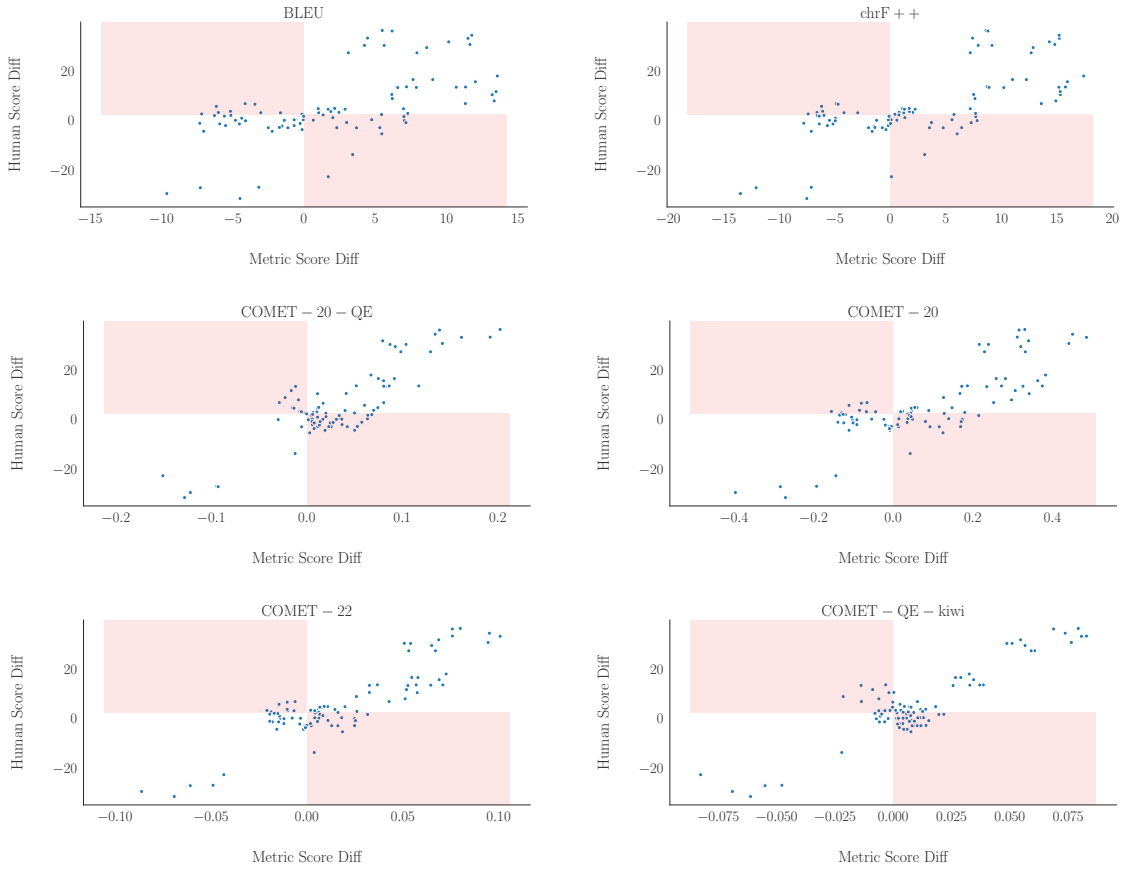


Figure 2: Pairwise comparison plots for existing metrics.

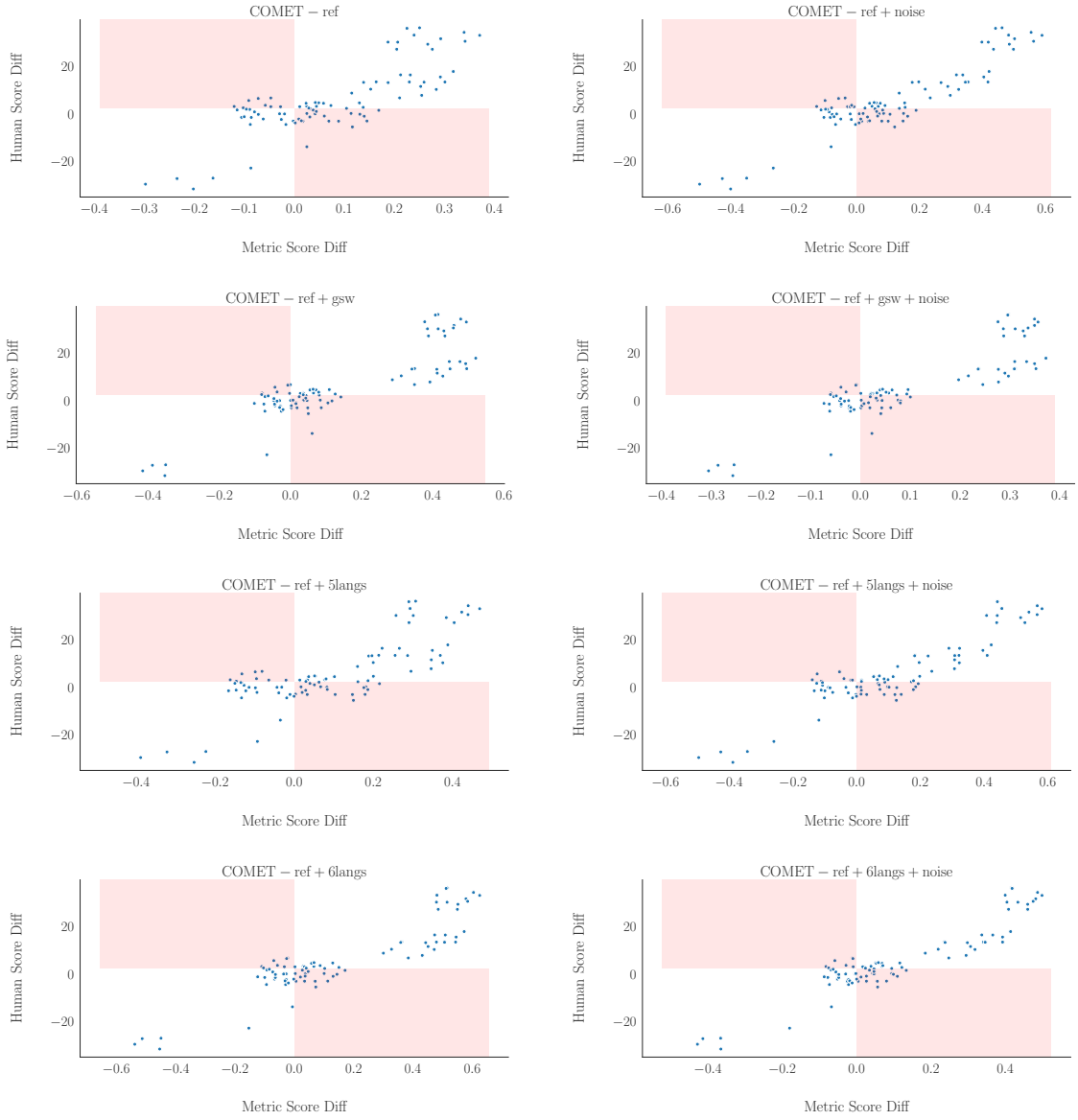


Figure 3: Pairwise comparison plots for the COMET-ref metrics trained for this work.

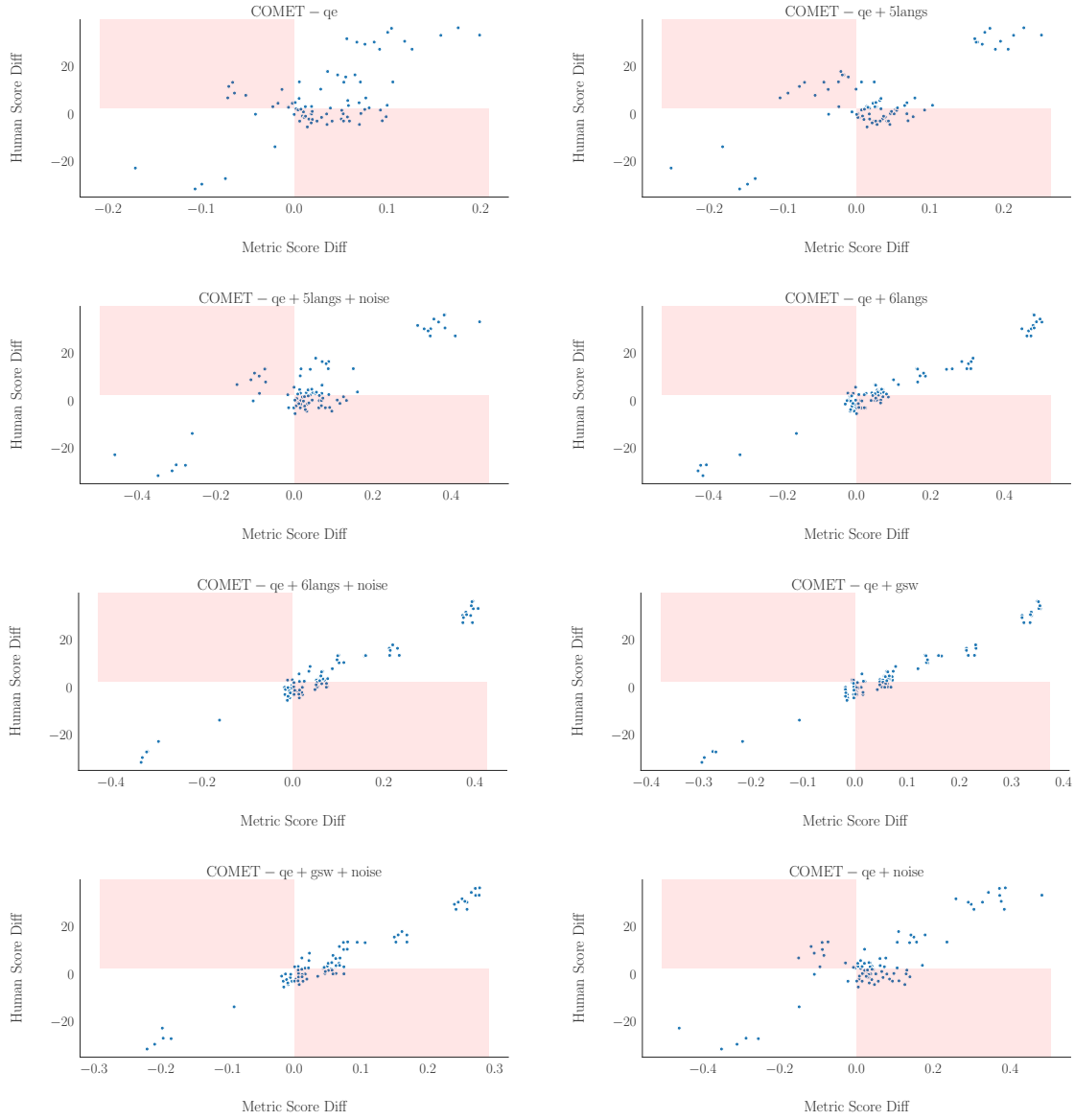


Figure 4: Pairwise comparison plots for the COMET-qe metrics trained for this work.

The Devil is in the Errors: Leveraging Large Language Models for Fine-grained Machine Translation Evaluation

Patrick Fernandes[*,2,3,4] **Daniel Deutsch**[1] **Mara Finkelstein**[1] **Parker Riley**[1]
André F. T. Martins[3,4,5] **Graham Neubig**[2,6]
Ankush Garg[1] **Jonathan H. Clark**[1] **Markus Freitag**[1] **Orhan Firat**[1]
[1]Google [2]Carnegie Mellon University [3]Instituto Superior Técnico
[4]Instituto de Telecomunicações [5]Unbabel [6]Inspired Cognition
pfernand@cs.cmu.edu

Abstract

Automatic evaluation of machine translation (MT) is a critical tool driving the rapid iterative development of MT systems. While considerable progress has been made on estimating a single scalar quality score, current metrics lack the informativeness of more detailed schemes that annotate individual errors, such as Multidimensional Quality Metrics (MQM). In this paper, we help fill this gap by proposing **AUTOMQM**, a prompting technique which leverages the *reasoning* and *in-context learning* capabilities of large language models (LLMs) and asks them to identify and categorize errors in translations. We start by evaluating recent LLMs, such as PaLM and PaLM-2, through simple *score prediction* prompting, and we study the impact of labeled data through in-context learning and finetuning. We then evaluate AUTOMQM with PaLM-2 models, and we find that it improves performance compared to just prompting for scores (with particularly large gains for larger models) while providing interpretability through error spans that align with human annotations.

1 Introduction

Evaluating natural language generation systems has always been challenging, and as the output quality of these systems has improved, evaluation has become even more challenging and critical. For example, in Machine Translation (MT), a field where evaluation has garnered considerable attention, previous standard automatic surface-level metrics such as BLEU (Papineni et al., 2002) are becoming less reliable as the quality of generation systems improves, with little remaining correlation with human judgments (Freitag et al., 2022).

To keep pace with the constantly improving quality of MT output, the next generation of automatic metrics is rapidly evolving. *Learned* automatic metrics that leverage human-judgments to finetune

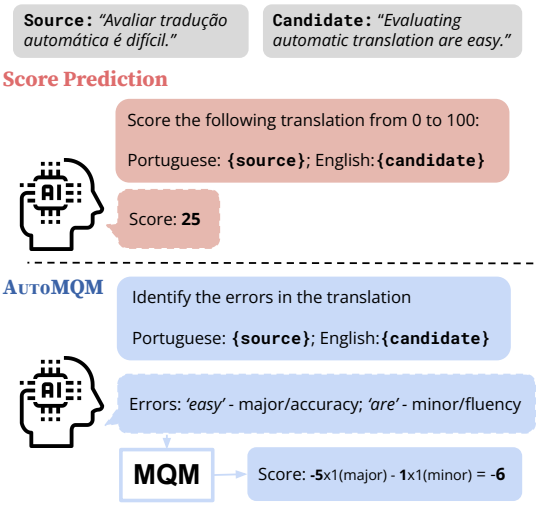


Figure 1: Illustration of how AUTOMQM uses LLMs to assess the quality of a translation. Rather than asking for a single quality score, AUTOMQM prompts models to identify and classify errors, and uses the MQM framework to produce a score.

language models (Sellam et al., 2020; Rei et al., 2022a) currently represent the state-of-the-art in automatic evaluation benchmarks like the WMT Metrics task (Freitag et al., 2022), and show high correlation with human judgments. However, these metrics typically output a single, *uninterpretable* quality score, making it difficult to understand the type and extent of errors identified by them. The lack of insights makes it difficult for model developers to leverage these metrics to improve their systems.

Unlike automatic metrics that only provide a single scalar value as quality score, state-of-the-art human evaluation methodologies like Multidimensional Quality Metrics (MQM; Lommel et al., 2014; Freitag et al., 2021a) ask professional annotators to identify and label error spans with a category and severity. This much richer feedback can be used to gain a better understanding of the current limitations of the model under evaluation and improve it.

In this paper, we ask whether large language

[*] Work done while working part-time at Google.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 1066–1083
December 6–7, 2023.

models (LLMs) in combination with a few human annotations can be used to design an automatic metric that generates rich feedback similar to that generated by human experts in MQM. This work is motivated by recent papers that demonstrated that LLMs can be used as automatic metrics (Liu et al., 2023b) to generate a single quality score. In particular, Kocmi and Federmann (2023) showed that LLMs can be prompted to assess the quality of machine-generated translations, even achieving state-of-the-art performance on assessing system-level quality. However, previous work only provides a limited view of the capabilities of LLMs for machine translation evaluation: the focus has predominantly been on *score prediction* (i.e. predicting a numerical value for quality), without considering the use of *any* annotated data (either through in-context learning or finetuning), and only in *high-resource* language pairs.

We provide a large-scale study of the capabilities of LLMs (from the PaLM and PaLM-2 families; Chowdhery et al., 2022; Anil et al., 2023) for machine translation evaluation (both with and without a reference translation), provide a novel comparison between prompting and finetuning, and investigate the performance in the low-resource scenario. Inspired by findings that the performance of LLMs can be improved by prompting them for *rationales* of their predictions (Wei et al., 2022; Lu et al., 2023), we also propose **AUTOMQM**, a prompting technique for MT evaluation that asks LLMs to identify error spans in a translation and to classify these errors according to the MQM framework, with a quality score derived automatically from the identified errors. A key advantage of AUTOMQM is its *interpretability*, as users can inspect the errors responsible for a score (Figure 1).

Our contributions can be summarized as follows:

- We confirm the finding of Kocmi and Federmann (2023) that LLMs are *zero-shot* state-of-the-art system-level evaluators, but show low correlation with human judgment compared to *learned* metrics at the segment-level.
- We show that *finetuning* an LLM with human judgment mitigates its low segment-level performance (particularly for smaller LLMs), showing similar correlations with human judgment at both the system-level and segment-level to state-of-the-art learned metrics.
- We are the first to evaluate LLM-based evaluation methods on low-resource language pairs. We find that their performance is promising, but lags behind state-of-the-art learned metrics.
- We find that, with AUTOMQM, PaLM-2 models can be prompted to generate rich MQM-like annotations, outperforming their score prediction counterparts at the segment-level.
- Furthermore, annotations predicted by PaLM-2 models correctly identify over 50% of words that are part of *major* errors, and are comparable to the ones produced by state-of-the-art *supervised* word-level evaluators.

Our findings might have significant implications for not only MT evaluation, but evaluation of machine-generated text in general, and further highlight the potential of using LLMs to provide *AI Feedback* (Fernandes et al., 2023).

The outputs of our models prompted with AUTOMQM are available at github.com/google-research/google-research

2 Background: MT Evaluation

Machine translation evaluation is one of the most well-studied evaluation problems in NLP (Callison-Burch et al., 2008; Freitag et al., 2022). In this task, given

1. a *source* sentence in a (source) language
2. a *candidate* translation in a (target) language

an evaluation metric assesses the quality of the candidate translation by how well it conveys the meaning of the source sentence while considering other factors like *fluency*. Like many other natural language generation evaluation problems, this task is difficult because the set of correct translations for a given source sentence is often very large and not entirely known in advance. To simplify the problem of machine translation evaluation, often (3) a *reference* translation (typically created by a professional human translator) is included as additional information when assessing the candidate translation. This sub-problem is known as *reference-based* evaluation (as opposed *reference-less* evaluation or *quality estimation*).

Up until recently, human evaluation of machine translation was carried out predominantly with the aim of assigning a single quality score to a candidate translation. Consequently, *learned* metrics, which leverage collected human judgment data, are trained for and evaluated on the same task of *score*

prediction (i.e., assigning a single quality score to a candidate translation), and can achieve high correlation with human-provided scores (Freitag et al., 2022).

However, framing machine translation evaluation as a score prediction task is problematic: any scoring or ranking of translations is implicitly based on an identification of errors in the candidate translations, and asking raters to solely provide a single score can lead to rushed and noisy judgments (Freitag et al., 2021a).

This insight has led to the adoption of the Multidimensional Quality Metrics (MQM) framework (Lommel et al., 2014; Freitag et al., 2021a) as the gold standard for evaluating machine translation. The MQM framework asks human evaluators to identify error spans in candidate translations and classify those errors according to various dimensions, e.g., *fluency*, *accuracy*, ... (see Appendix A for a more detailed description of MQM). Importantly, the MQM framework *does not* ask annotators to provide a quality score for each translation, and instead derives one automatically from the identified error spans and their classifications. However, despite its richness, *most* automatic metrics that leverage MQM data only use the final quality score produced by the framework and discard the error span information and classification.

3 Related Work

The success of *learned* machine translation metrics (Sellam et al., 2020; Rei et al., 2022a; Freitag et al., 2022; Qin et al., 2022), which finetune neural network models pretrained on large amounts of (unsupervised) data, highlighted the importance of leveraging *transfer learning* to achieve metrics with better correlation with human judgments. More recently, *generative* LLMs (OpenAI, 2023; Anil et al., 2023) have consistently demonstrated impressive results in natural language understanding and *zero-* and *few-shot* transfer and, naturally, interest in employing these models for (translation) evaluation has increased. Kocmi and Federmann (2023) first explored the use of GPT models for evaluating machine translation tasks, showing their potential as *zero-shot* evaluators, and others have since extended GPT-based evaluation to other generation problems (Jain et al., 2023; Liu et al., 2023b).

Perrella et al. (2022) first highlighted that MQM annotations could be leveraged to allow pretrained models to predict major and minor errors and, similarly to AUTOMQM, used the identified errors to automatically score translations. However, their approach relied on weaker encoder-only or encoder-decoder language models, required *supervised* data to work, and overall underperformed other top metrics. We compare against their *MaTASe* metric in our experiments. Lu et al. (2023) showed that doing *error analysis*, a prompting technique similar to AUTOMQM, could lead to better ChatGPT-based evaluators. However, they still relied on the LLM to provide a score once it identified errors (rather than do it automatically using something like the MQM framework). Furthermore, they provided a very limited meta-evaluation using only 40 examples per language pair. Concurrently with our work, Xu et al. (2023) proposed INSTRUCTSCORE, a LLaMA-based evaluator that asks models to identify and categorize errors in translation (as well as providing a natural language explanation for each error). However, the authors only explore a 7B parameter model and don't leverage zero- and few-shot capabilities of models as in this work. Instead, they rely on a more complex approach of distilling the knowledge of a more capable GPT-4 LLM.

Additionally, WMT Word-Level Quality Estimation shared tasks (Fonseca et al., 2019; Zerva et al., 2022) leverage MQM data by converting span-level annotations of errors (normally of *major* severity) to word-level tags and Task 2 in the WMT19 Quality Estimation shared task evaluation explicitly evaluated submissions of span-level annotations (although most submissions still consisted of models that predicted word-level tags which were converted to spans). We also compare against state-of-the-art word-level quality estimation models.

4 Using LLMs to Predict Quality Scores

Recent works have shown that large language models are versatile, general-purpose models that can be used to tackle many problems in NLP, including evaluation (Kocmi and Federmann, 2023; Jain et al., 2023; Liu et al., 2023b). We begin by exploring how LLMs can be used for machine translation evaluation through *score prediction*.

4.1 Prompting

We start by measuring how far we can push the performance of LLMs with just *prompting* (Liu et al., 2023a): by defining the task of MT evaluation and quality estimation as *textual templates* (with

a general description of the problem and "slots" for the inputs and outputs), we can use general-purpose LLMs to perform these tasks at inference-time, without any parameter updates.

Throughout the paper, we choose to use Kocmi and Federmann (2023)'s `GEMBA-SQM` prompt (Figure 9, Appendix C), which asks models to generate (a string representation of) a score from 0-100. We choose this prompt for two reasons: firstly, early explorations with various prompts showed that this generally performed well. Secondly, using a single prompt ensures a fairer comparison between the capabilities of different models.[1]

In-Context Learning A surprising emergent capability of LLMs is their ability to improve on prompting-based tasks by including a very small amount of labeled data as part of the prompt/context (Brown et al., 2020) and *without* parameter updates, a technique called *in-context learning* (ICL) or *few-shot prompting*. We thus investigate the impact that ICL has on LLMs' ability to assess translation quality. Recent works have shown that the impact of ICL is tightly tied with the exact examples included in the prompt, with a poor selection procedure leading to no improvements or even worse performance than the zero-shot case (Jain et al., 2023). We therefore explore two sampling approaches to select in-context examples from a pre-defined "pool" of translation quality assessments: **uniform** and **stratified sampling**, where the example pool is bucketed by score ranges and examples are sampled from each bucket.

4.2 Finetuning

It has previously been shown that LLMs are capable of zero-shot evaluation (Kocmi and Federmann, 2023), but the extent to which *finetuning* on human judgment data can further boost the performance of LLMs has not been studied. In the WMT'22 Metrics Shared Task (Freitag et al., 2022), all top submissions were learned metrics; that is, pretrained models finetuned on human judgment data[2].

Thus, we investigate whether LLMs are amenable to finetuning on human judgment data. LLMs used in top-performing metrics are generally much larger than the pretrained language models leveraged by previous learned metrics (which generally have fewer than 1 billion parameters). Moreover, most learned metrics leverage pretrained encoder-only rather than (decoder-only) prefix language models. We experiment with finetuning LLMs using two objectives:

- *Regression* (**R**): Commonly used for training learned metrics (Rei et al., 2022a), the objective here is a regression loss (e.g., mean squared error) between continuous scores obtained from the model (for example, with a *regression head*) and the human scores.
- *Generative Classification* (**GC**): We bucket scores into discrete classes (e.g. *"bad"*, *"ok"* and *"good"*) and treat the MT evaluation task as a text-to-text classification problem (Raffel et al., 2020) by having the model generate a template sentence with the class. See §6.1 for more details.

5 Using LLMs to Predict Error Spans

While producing quality scores that correlate with human judgments is an important part of translation quality assessment, metrics that solely do score prediction suffer from problems of **interpretability**: if a metric assigns a low score, the downstream users are left in the dark about which parts of the translation were responsible for the score and thus need to be corrected. This is especially problematic in cases where the metric assigns a *wrong* score to a translation, as it is much harder to diagnose why the evaluation model made a mistake, and identify and prevent similar mistakes in the future. In fact, reducing translation quality to a single score has proven problematic even for human annotators: asking raters to solely provide a single score can lead to rushed and noisy judgments (Freitag et al., 2021a) and the current gold standard for translation quality evaluation involving human annotators is instead based on methodologies like the MQM framework (see §2) , which provide richer feedback by identifying error spans, categorizing them, and evaluating their severity.

Interestingly, another emergent phenomenon in LLMs is the success of *chain-of-thought* prompting (Wei et al., 2022): when defining a prompt for a particular task, if we instruct the model to produce a series of intermediate reasoning steps (*"let's think step-by-step"*), it tends to generate a free-text *rationale* before generating an output, and this often improves the performance on the

[1] While this prompt wasn't the best for *system-level*, it led to the best *segment-level* performance in GEMBA.

[2] While these metrics all leverage powerful pretrained (language) models, these generally aren't considered LLMs

```
Based on the given source and reference, identify the major and minor errors in this
translation. Note that Major errors refer to actual translation or grammatical errors,
and Minor errors refer to smaller imperfections, and purely subjective opinions about
the translation.

{src_lang} source: "{source}"
{tgt_lang} human reference: "{reference}"
{tgt_lang} translation: "{candidate}"
Errors: {error1:span} - {error1:severity}/{error1:category}; {error2:span} - ...
```

Figure 2: The AUTOMQM prompt used in this paper. Parts in purple are only included for *reference-based* evaluation, while parts in orange represent slots for outputs, and are only included for in-context examples.

task at hand (Liu et al., 2023b). Furthermore, this *chain-of-thought* prompting can be used to obtain *structured* rationales from LLMs, and this can lead to better performance than with free-text rationales (Lu et al., 2023).

Motivated by these findings, we propose **AUTOMQM**, a prompting technique for translation quality assessment that instructs LLMs to *identify* errors in a translation, and *categorize* the type of error according to the MQM framework (Lommel et al., 2014). Furthermore, we *don't* ask the model to produce a score, as the MQM framework provides an algorithmic procedure to obtain one from identified errors: the total score is the sum of penalties for all errors identified, where (roughly) *major* errors get penalized with -5 and *minors* with -1 (see Appendix A for a more detailed description of the scoring algorithm).[3] Figure 2 shows the main AUTOMQM prompt used in this paper.

Importantly, obtaining meaningful AUTOMQM results in a zero-shot setting is a substantially more challenging task compared to score prediction: we found that, without any in-context examples, LLMs tend to produce outputs that are either uninformative or difficult to parse. Thus we only consider the AUTOMQM task in the *few-shot* scenario. Based on the findings from §6.2, we explore the impact of in-context learning by sampling from the example pool using stratified sampling extended with a set of *rejection criteria* (Appendix D), which ensures that the example set has a balance between major and minor errors as well as diversity in the categories of errors.

6 Experiments

6.1 Experimental Setup

Data The metrics in this work are evaluated on both *high-resource* and *low-resource* language pairs. The three high-resource language pairs come from the WMT'22 Metrics Shared Task (Freitag et al., 2022): en→de, zh→en, and en→ru. The ground-truth translation quality scores are derived from MQM ratings in which expert annotators marked error spans in the translations with different severity levels which are automatically converted to a numeric score (see §2). The four low-resource language pairs come from the WMT'19 Metrics Shared Task (Ma et al., 2019): en↔gu and en↔kk. Since MQM ratings are not available for the low-resource pairs, the ground truth quality scores are direct assessment (DA) scores. DA scores are quality assessments assigned by non-expert raters on a scale from 0-100, normalized per rater. See Table 9 (Appendix B) for statistics about the number of MT systems and segments for every language pair.

Additionally, in our experiments, AUTOMQM required in-context examples with MQM annotations to work, so we restrict our evaluation of AUTOMQM to en→de and zh→en because there are available MQM ratings from the WMT'21 Metrics Shared Task (Freitag et al., 2021b) that we can use as in-context learning example pools.

Models We base most of our experiments on the following LLMs:

- **PaLM**: A 540 billion parameter autoregressive Transformer model trained on 780 billion tokens of high-quality text (Chowdhery et al., 2022). It showed remarkable performance on a wide-range of NLP tasks, including Machine Translation (Vilar et al., 2022).
- **PaLM-2**: The successor to PaLM, the PaLM-2 family of LLMs (Anil et al., 2023) builds upon recent research insights, such as compute-optimal scaling, a more multilingual and diverse pre-training mixture, and architectural/optimization improvements. We mainly use two model sizes in the family: PaLM-2 BI-

[3] This is similar to methods that leverage external *executors* to improve the performance of LLMs (Gao et al., 2022)

SON and (the larger) PaLM-2-UNICORN.[4] In addition we explore the impact of instruction-tuning by using a UNICORN model finetuned on the FLAN dataset (Wei et al., 2021).

For *score prediction*, we compare PaLM and PaLM-2 against the GPT family of LLMs (Brown et al., 2020; OpenAI, 2023) by leveraging the results and outputs from the GEMBA evaluator (Kocmi and Federmann, 2023). We then evaluate the performance of AUTOMQM with only PaLM-2 models (which performed best in score prediction).

Additionally, for the high-resource languages, we compare to a set of strong baseline evaluation metrics, MetricX-XXL and COMET-22, which were the two top-performing metrics in the WMT'22 Metrics Shared Task. MetricX-XXL and COMET-22 are both finetuned regression models trained on DA data from WMT that are initialized with mT5 (Xue et al., 2021) and XLM-R (Conneau et al., 2020), respectively.

For the AUTOMQM experiments, we also compare against MATESE, a comparable submission to the WMT'22 Metrics Shared task that finetuned a XLM-R model to identify major and minor errors, and computed a score automatically. Since we were unable to obtain the span-level predictions for the MATESE submission, we also compare against the top submission to the WMT'22 Word-Level Quality Estimation Shared Task (Zerva et al., 2021): word-level COMETKIWI (COMET-WL) (Rei et al., 2022b), also based on an XLM-R model trained on a combination of sentence- and word-level data. To do so, we re-run this model on the WMT'22 Metrics Shared Task data, and convert the predicted *word-level* `OK/BAD` tags into spans.[5]

Finetuning For *regression* finetuning, we use a real-valued logit, extracted from a fixed index in the first target token's logit vector, as the quality signal. (In particular, we leverage a special, *unused*, vocabulary token.) This was the technique used to train MetricX-XXL in the WMT 2022 Shared Task submission (Freitag et al., 2022). The regression-based model was trained on WMT direct assessment (DA) data from the years 2015 through 2020.

For *generative* classification, we bucket the scores in the training data into five classes, where class boundaries are assigned so that each class contains an equal number of training examples. We then map labels to verbal ratings from the following set, based on their bucket: [*"very bad"*, *"bad"*, *"ok"*, *"good"*, *"very good"*]. To evaluate the model, predictions are mapped back to integer labels from 1 to 5. Any predictions not containing a substring in the label set are considered invalid and are mapped to 0. We experimented with finetuning on both DA and MQM 2020 (Freitag et al., 2021a) data, and found that the latter performed slightly better.

To assess the impact of *model size*, we also finetune two additional (smaller) PaLM-2 models, which we call S and M, comparing their finetuned and zero-shot performance.[6]

Metric Meta-Evaluation The quality of an automatic evaluation metric is estimated by comparing the agreement between the metric scores and ground-truth quality scores on a large number of translations from different MT systems, a process known as metric meta-evaluation. This work reports three different agreement scores, as follows.

The first is system-level accuracy, which calculates the percent of system pairs that are ranked the same by the metric and ground-truth scores, micro-averaged over a set of language pairs (Kocmi et al., 2021). System-level scores are defined as the average score across all segments.

At the segment-level, the standard correlation that is reported by WMT is Kendall's τ. However, recent work pointed out problems with Kendall's τ with respect to ties (Deutsch et al., 2023). In short, different variants of τ are inconsistent with respect to ties and even biased against metrics that predict ties, as our metrics do in this work. Deutsch et al. (2023) recommend reporting a pairwise accuracy score, which rewards metrics for correctly ranking translations as well as correctly predicting ties, in combination with a tie calibration procedure that automatically introduces ties into metric scores so that the meta-evaluation is fairer. This accuracy score, denoted acc*, ranges between 0 and 1, and a random metric would achieve 33% accuracy. We report the "group-by-item" variant of the pairwise accuracy score from Deutsch et al. (2023) in addition to Pearson's ρ, a complementary signal to rank-based correlations that measure the strength of the linear relationship between two variables (and one of the standard correlations reported in WMT).

[4] Information about exact number of parameters of PaLM-2 models is not publicly available.

[5] We consider a span as any maximal consecutive sequence of words marked as `BAD`, assigning every span the *major* severity.

[6] We use a small variation of the *zero-shot* prompt, asking models for scores from the same 5 buckets used in finetuning.

		System-Level	Segment-Level					
		All (3 LPs)	EN-DE		ZH-EN		EN-RU	
Model	Ref?	Accuracy	ρ	acc*	ρ	acc*	ρ	acc*
Baselines								
MetricX-XXL	✓	85.0%	0.549	61.1%	0.581	54.6%	0.495	60.6%
COMET-22	✓	83.9%	0.512	60.2%	0.585	54.1%	0.469	57.7%
COMET-QE	✗	78.1%	0.419	56.3%	0.505	48.8%	0.439	53.4%
Prompting								
PaLM 540B	✓	90.1%	0.247	55.4%	0.255	48.5%	0.180	48.6%
PaLM-2 BISON	✓	88.7%	0.394	56.8%	0.322	49.3%	0.322	52.8%
PaLM-2 UNICORN	✓	90.1%	0.401	56.3%	0.349	51.1%	0.352	55.3%
FLAN-PaLM-2 UNICORN	✓	75.9%	0.197	55.6%	0.139	46.1%	0.198	52.0%
PaLM 540B	✗	84.3%	0.239	56.1%	0.270	43.1%	0.300	51.8%
PaLM-2 BISON	✗	85.0%	0.355	57.0%	0.299	48.6%	0.303	53.1%
PaLM-2 UNICORN	✗	84.3%	0.275	56.1%	0.252	48.3%	0.209	49.8%
FLAN-PaLM-2 UNICORN	✗	69.7%	0.116	54.6%	0.112	43.8%	0.156	47.8%
Finetune								
PaLM-2 BISON (**R**)	✓	88.0%	0.511	61.0%	0.459	51.5%	0.458	59.5%
PaLM-2 BISON (**GC**)	✓	86.1%	0.400	59.2%	0.444	49.3%	0.365	56.0%
PaLM-2 UNICORN (**R**)	✓	87.6%	0.508	61.1%	0.412	52.6%	0.460	60.4%
PaLM 2 BISON (**R**)	✗	87.6%	0.490	59.9%	0.439	53.4%	0.437	59.2%
PaLM 2 BISON (**GC**)	✗	86.1%	0.368	57.5%	0.420	47.3%	0.390	54.9%
PaLM 2 UNICORN (**GC**)	✗	86.1%	0.407	57.9%	0.402	45.6%	0.411	55.3%

Table 1: Meta-evaluation results at system and segment-level for the *high-resource* language pairs. Finetuned **(R)** and **(GC)** represent the *regression* and *generative classification* objectives (§4.2). ✓and ✗ represent *reference-based* and *reference-less* metrics, respectively.

Span Meta-Evaluation Since AUTOMQM provides not only scores but also the identified error spans, we can compare the predicted spans with the errors marked by annotators in the MQM annotations. We evaluate quality of predicted spans using: (1) *Span Precision* (SP), which measures the overlap of predicted spans and gold (annotated) spans; and (2) *Major recall* (MR), which captures the percentage of gold major errors that were predicted as errors (either minor or major).

More formally, consider the set of ground truth spans $S^\star$, where each span consists of a sequence of words, i.e., $s_i = (w_{(a)}, w_{(a+1)}, \cdots)$. Let $S^\star_{\text{maj}} \subseteq S^\star$ be the subset containing only the major errors. Given a span set S, we define its positional set $P(S)$ as the set containing the positions of all the words in every span in S. For example, assuming a span $s_i = (w_{(n)}, w_{(n+1)}, \cdots)$ in S starts at the nth position in the text, its corresponding positional set will include the positions $\{n, n+1, ..., n+\text{len}(s_i)-1\}$. Then for a set of *predicted* spans $\hat{S}$, SP and MR are defined as:

$$\text{SP}(\hat{S}) = \frac{|P(\hat{S}) \cap P(S^\star)|}{|P(\hat{S})|} \quad (1)$$

$$\text{MR}(\hat{S}) = \frac{|P(\hat{S}) \cap P(S^\star_{\text{maj}})|}{|P(S^\star_{\text{maj}})|} \quad (2)$$

Intuitively, we care for overall precision (regardless of severity) since we want to make sure predicted errors tend to be marked by annotators as well, but for recall we care mostly for *major* errors, as these have a larger impact on translation quality and are more critical to identify. Additionally, we also report the (3) *Matthews Correlation Coefficient* (MCC), one of the official metrics in the word-level quality estimation tasks (Zerva et al., 2022).

6.2 Results

6.2.1 Score Prediction

Table 1 summarizes the meta-evaluation results, at the *system* and *segment* level, for both the *zero-shot prompting* and *finetuning* settings.

Prompting A first observation is almost all zero-shot LLM evaluators have higher *system-level* performance than learned metrics (with and without references), with PaLM 540B and PaLM-2 UNICORN achieving the best performance. At the segment level, the story is more complicated: similarly to Kocmi et al. (2022), we find that none of the LLMs we explored was able to consistently outperform the baseline learned metrics. We see that PaLM-540B is a particularly poor reference-based evaluator, which is surprising given its system-level performance. Unexpectedly, instruction-tuning with FLAN seems to *degrade* performance, with FLAN-PaLM-2 UNICORN achieving poor performance at both the system and segment levels.[7]

Nevertheless, PaLM-2 models achieve high correlations with human judgments, and the *reference-*

[7]Note that this might be a problem with the FLAN dataset and not instruction-tuning in general, as the GPT models are also instruction-tuned and perform well.

		System	Segment acc*		
Model	Ref?	All	EN-DE	ZH-EN	EN-RU
GEMBA					
GPT-3.5	✓	85.4%	54.9%	49.5%	47.5%
GPT-4	✓	88.7%	57.8%	52.6%	55.0%
GPT-3.5	✗	82.5%	56.1%	49.7%	49.3%
GPT-4	✗	89.1%	56.4%	53.4%	54.8%
BISON	✓	88.7%	56.8%	49.3%	52.8%
UNICORN	✓	90.1%	56.3%	51.1%	55.3%
BISON	✗	85.0%	57.0%	48.6%	53.1%
UNICORN	✗	84.3%	56.1%	48.3%	49.8%

Table 2: Comparison between PaLM-2 and GPT-based GEMBA (Kocmi et al., 2022) at the system and segment levels for the *high-resource* language pairs.

less PaLM-2 BISON is competitive with the *learned* baselines, particularly at assessing alternative translations of the same sentence (acc*). When comparing PaLM-2 models with Kocmi et al. (2022)'s GPT-based GEMBA evaluator (Table 2), we see that both families of LLMs perform similarly, with PaLM-2 models exhibiting higher system-level performance than GPT-based GEMBA, while GEMBA achieves better segment-level accuracy, particularly in the reference-less setting.

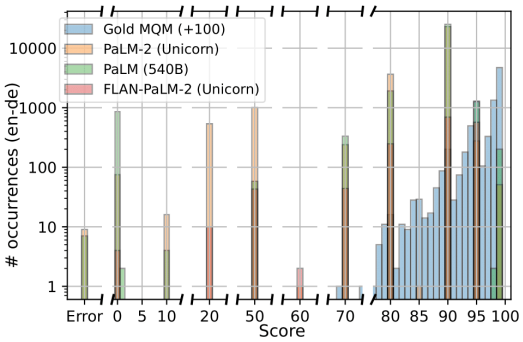


Figure 3: Distribution of scores for various LLM *reference-based* evaluators, on the EN-DE test set. Note that the y axis is in *log-scale*.

Figure 3 shows the distribution of scores produced by PaLM- and PaLM-2-based evaluators. We find that, despite being prompted to give a score in the 0-100 range, these models almost always output one of a very limited set of scores (e.g. 0, 50, 90, 95). Given Kocmi and Federmann (2023)'s similar findings with GPT models, it seems that this is a consequence of the pretraining objective.

Finetuning Despite their already-great performance in the zero-shot setting, we find that finetuning LLMs can further improve LLM evaluators' segment-level scores. This is particularly obvious for the *reference-less* evaluators, where a finetuned PaLM-2 BISON achieves state-of-the-art performance in segment-level correlations and comparable system-level accuracy across all language pairs. Moreover, when we look at how performance *scales* with parameter count (Figure 4), we observe an interesting trend: while smaller models are not capable of being effective zero-shot evaluators, finetuning them leads to competitive performance, and only a slight decrease when compared to their larger finetuned counterparts.

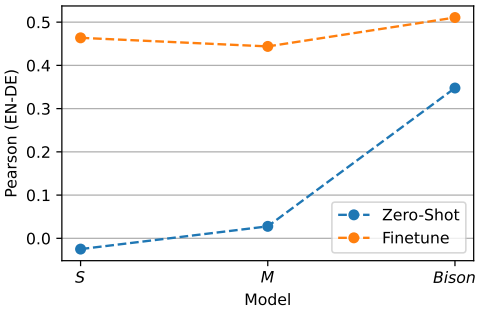


Figure 4: Behavior of *Pearson* as we scale the LLM's parameter count. Note that the x axis is not to-scale with regard to parameter count.

In-context Learning Figure 5 shows the mean and interquartile range (IQR) of the performance as we increase the number of in-context examples k (with 100 example sets per k) sampled with *stratified* sampling (see Appendix E for *uniform*). Surprisingly, despite evidence of the benefits of in-context learning for many tasks, we found that including in-context examples during evaluation (almost) never led to better performance, either with *uniform* or *stratified* sampling.

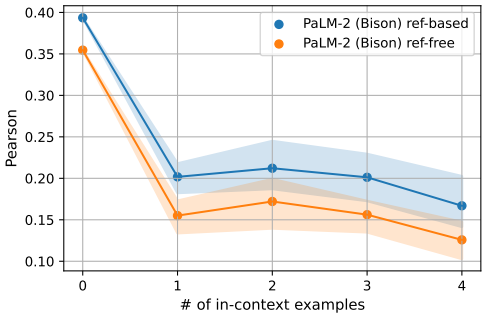


Figure 5: Mean *Pearson* and its interquartile range (IQR) in the WMT22 EN-DE test set, as we increase the number of in-context examples with *stratified* sampling

To investigate the cause of this disappointing performance, we looked at how *particular* in-context example sets affect the distribution of scores produced by LLM-based evaluators. Figure 6 shows the distribution of scores *over the whole test set* for the 1-shot and 2-shot settings, with different in-context examples sets. We can see that output distribution is heavily biased by the scores in the in-context examples: despite *never* predicting 79

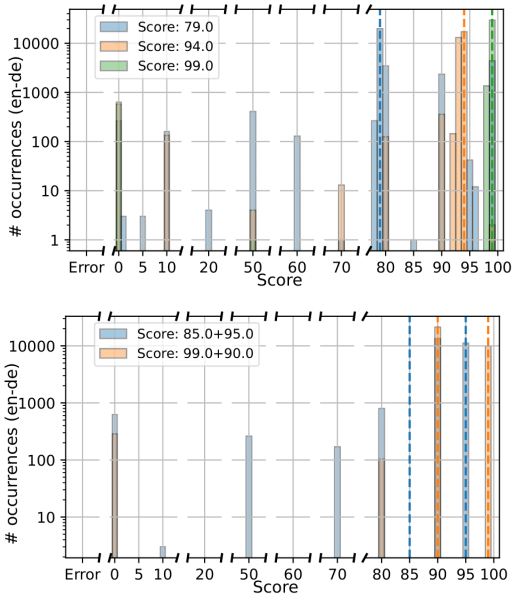


Figure 6: Distribution of scores for PaLM-2 (BISON) models for 1-shot (top) and 2-shot (bottom) setups, with various in-context learning sets for each (and their scores in the legend)

in the zero-shot setting, when a single example with that score is included, it starts to dominate the model predictions. This seems to hint that LLMs "overfit" to the specific scores provided as examples, rather than generalizing to the broader evaluation task, which could explain the lackluster performance of in-context learning.

6.3 Low Resource Languages

Table 3 shows the performance of PaLM-2 models at *score prediction* for *low-resource* translation. Overall, we find that similar to high-resource LPs, these models are good zero-shot evaluators, with system-level accuracies around 90%. However, *zero-shot* LLMs underperform *learned* metrics, even when these metrics also weren't exposed to data in these low-resource languages.

		System	Segment ρ			
Model	Ref?	All	EN-KK	EN-GU	KK-EN	GU-EN
Baseline						
MetricX-XXL*	✓	94.0%	0.666	0.701	0.539	0.409
Prompting						
BISON	✓	92.2%	0.605	0.540	0.462	0.339
UNICORN	✓	87.4%	0.609	0.621	0.495	0.384
BISON	✗	89.8%	0.567	0.478	0.381	0.313
UNICORN	✗	84.4%	0.536	0.523	0.433	0.334

Table 3: Meta-evaluation results for system-level *accuracy* and segment-level *Pearson* on the low-resource languages, using PaLM-2 for *score prediction*. *Note that the baseline is slightly different from the high-resource case, being trained on the same data but *without* these *low-resource* language pairs.

6.3.1 AUTOMQM

Figure 14 shows the mean and interquartile range (IQR) of the performance of PaLM-2 BISON with AUTOMQM, as we increase the number of in-context examples (again, with 100 example sets per k). Contrary to the performance with score prediction, we find that performance with AUTOMQM seems to (mostly) scale with the number of in-context examples: performance increases monotonically with up to 4 in-context examples and plateaus thereafter. Additionally, the variance across the in-context learning sets seems to be lower, with most example sets exhibiting less than 0.05 *Pearson* difference from the best-performing sets. All this suggests that LLM evaluators are much more robust to the choice of in-context examples when prompted for AUTOMQM rather than for score prediction. We also find that the behavior of in-context learning is quite similar for both reference-based and reference-less evaluation tasks. Finally, we observe that the example sets that perform well for one task generally work well for the other, with performance on both settings given a fixed in-context set being highly correlated, as shown in Figure 7.

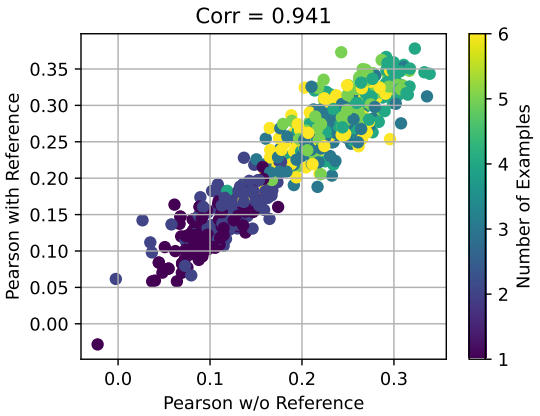


Figure 7: Scatter plot of the *Pearson* of PaLM-2 (BISON) models, with/without including the *reference* in the prompt, for each in-context learning setting tried.

Table 4 shows the meta-evaluation results for PaLM-2 BISON and UNICORN prompted with AUTOMQM (using the best-performing in-context learning sets in Figure 14). For ease of comparison, we also report their performance when prompted for *score prediction*, as well as the performance of the baselines. Overall, prompting LLMs with AUTOMQM seems to lead to significant improvements in evaluating machine translation quality, particularly for larger models: UNICORN achieves better performance (across all meta evaluations) with it than when prompted for *score prediction*,

		System-Level	Segment-Level			
		All (2 LPs)	EN-DE		ZH-EN	
Model	Ref?	Accuracy	ρ	acc*	ρ	acc*
Baselines						
MetricX-XXL	✓	81.1%	0.549	61.1%	0.581	54.6%
MATESE	✓	79.9%	0.391	58.8%	0.528	51.5%
COMET-QE	✗	76.9%	0.419	56.3%	0.505	48.8%
MATESE-QE	✗	73.4%	0.298	57.9%	0.468	50.1%
COMET-WL	✗	71.6%	0.418	57.1%	0.406	51.5%
Score Prediction						
PaLM-2 BISON	✓	86.4%	0.394	56.8%	0.322	49.3%
PaLM-2 UNICORN	✓	86.4%	0.401	56.3%	0.349	51.1%
PaLM-2 BISON	✗	84.0%	0.355	57.0%	0.299	48.6%
PaLM-2 UNICORN	✗	80.5%	0.275	56.1%	0.252	48.3%
AutoMQM						
PaLM-2 BISON	✓	84.0%	0.369	59.2%	0.355	48.4%
PaLM-2 UNICORN	✓	87.6%	0.432	59.1%	0.442	51.8%
PaLM 2 BISON	✗	87.6%	0.297	55.2%	0.331	48.0%
PaLM 2 UNICORN	✗	83.4%	0.368	56.4%	0.429	50.2%

Table 4: Meta-evaluation results for PaLM-2 models using *AutoMQM* and score prediction, at the system and segment levels for multiple language pairs.

and its reference-less version is competitive with the best learned metric even at the segment level. However, for the smaller BISON, the benefits of AUTOMQM are less clear, with both techniques performing comparably. This hints that *scale* is necessary for *zero-* and *few-* shot fine-grained evaluation (like with AUTOMQM). We also find that the *distribution* of scores produced by LLMs prompted with AUTOMQM is much closer to the gold MQM distribution, with models outputting a much larger set of scores, and in the same ranges as annotators do (see Figure 8).

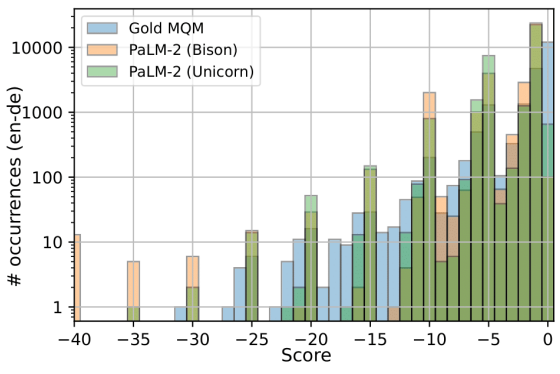


Figure 8: Distribution of scores for PaLM-2 models using AUTOMQM, on WMT22 EN-DE

Finally, when evaluating the error spans produced by LLMs prompted with AUTOMQM (Table 5), we find that PaLM-2 models are able to identify most of the *major* errors. However, it does seem to *over-predict* errors (with errors predicted by UNICORN having on average ∼5 words per span vs ∼2 words in the ground truth) and have overall low span precision. Similarly to overall *score* correlations, *scale* also seems to be important for the quality of spans produced by AUTOMQM, with UNICORN outperforming BISON at most metrics. Additionally, UNICORN prompted with AutoMQM predicts spans of comparable quality to the ones produced by current state-of-the-art *learned* word-level evaluators (trained on a considerable number of fine-grained annotations derived from MQM): while word-level models are more precise, their overall span correlation (MCC) is comparable, and they miss considerably more *major* errors than LLMs (despite only leveraging a handful of annotations).

		EN-DE			ZH-EN		
Model	R?	SP	MR	MCC	SP	MR	MCC
Baselines							
COMET-WL	✗	0.267	0.250	0.161	0.364	0.178	0.152
AutoMQM							
BISON	✓	0.095	0.749	0.060	0.252	0.255	0.109
UNICORN	✓	0.175	0.628	0.193	0.238	0.476	0.143
BISON	✗	0.119	0.520	0.092	0.224	0.311	0.091
UNICORN	✗	0.150	0.580	0.150	0.229	0.488	0.133

Table 5: Span-level meta-evaluation on WMT22 for PaLM-2 models using *AutoMQM*. **SR** and **MR** represent *span precision* and *major recall*, respectively.

7 Conclusion

In this study, we have systematically investigated the capabilities of large language models for machine translation evaluation through *score prediction*, and proposed AUTOMQM, a novel

prompting technique that leverages the Multidimensional Quality Metrics (MQM) framework for interpretable MT evaluation using LLMs.

We demonstrated that just prompting LLMs for score prediction leads to state-of-the-art system-level evaluators, but still falls short of the best *learned* metrics at the segment-level (with finetuning being necessary to close this gap). Then we showed that AUTOMQM can further improve the performance of LLMs without finetuning while providing interpretability through error spans that align with human annotations.

Our findings surrounding finetuning LLMs for *score prediction* hint that LLMs' performance in machine translation evaluation could be further improved by finetuning these models on fine-grained human judgment data (like MQM) and is a direction we are actively pursuing. Additionally, the general-purpose nature of LLMs may enable the application of similar prompting techniques (leveraging some fine-grained evaluation schemes) to other evaluation problems (Wu et al., 2023).

Acknowledgements

We would like to thank Ricardo Rei, Marcos Treviso and Chryssa Zerva for helping run the word-level QE baselines, and George Foster who provided feedback on an earlier version of this work. This work was partially supported by EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), by the project DECOLLAGE (ERC-2022-CoG 101088763), by the Portuguese Recovery and Resilience Plan through project C645008882- 00000055 (Center for Responsible AI), and the Fundação para a Ciência e Tecnologia through contracts SFRH/BD/150706/2020 and UIDB/50008/2020.

References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Josh Schroeder, and Cameron Shaw Fordyce, editors. 2008. *Proceedings of the Third Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Columbus, Ohio.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pil-

lai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties Matter: Modifying Kendall's Tau for Modern Metric Meta-Evaluation.

Patrick Fernandes, Aman Madaan, Emmy Liu, António Farinhas, Pedro Henrique Martins, Amanda Bertsch, José G. C. de Souza, Shuyan Zhou, Tongshuang Wu, Graham Neubig, and André F. T. Martins. 2023. Bridging the gap: A survey on integrating (human) feedback for natural language generation.

Erick Fonseca, Lisa Yankovskaya, AndrÃ© F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the wmt 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–12, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2022. Pal: Program-aided language models. *arXiv preprint arXiv:2211.10435*.

Sameer Jain, Vaishakh Keshava, Swarnashree Mysore Sathyendra, Patrick Fernandes, Pengfei Liu, Graham Neubig, and Chunting Zhou. 2023. Multi-dimensional evaluation of text summarization with in-context learning.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Revista Tradumàtica: tecnologies de la traducció*.

Qingyu Lu, Baopu Qiu, Liang Ding, Liping Xie, and Dacheng Tao. 2023. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt. *arXiv preprint*.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Niccolò Campolungo, and Roberto Navigli. 2022. MaTESe: Machine translation evaluation as a sequence tagging problem. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 569–577, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yiwei Qin, Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2022. T5score: Discriminative fine-tuning of generative evaluation metrics. *ArXiv*, abs/2212.05726.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2022. Prompting palm for translation: Assessing strategies and performance.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-grained human feedback gives better rewards for language model training.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Yang Wang, and Lei Li. 2023. Instructscore: Towards explainable text generation evaluation with automatic feedback.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Chrysoula Zerva, Daan van Stigt, Ricardo Rei, Ana C Farinha, Pedro Ramos, José G. C. de Souza, Taisiya Glushkova, Miguel Vera, Fabio Kepler, and André F. T. Martins. 2021. IST-unbabel 2021 submission for the quality estimation shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 961–972, Online. Association for Computational Linguistics.

A Multidimensional Quality Metric (MQM)

The Multidimensional Quality Metrics (MQM) framework is a flexible human-evaluation framework developed to evaluate and categorize errors in translations. Annotators are instructed to identify all errors within each segment in a document, paying particular attention to document context. See Table 6 for the annotator guidelines provided.

Annotators are asked to assign both an error *severity* and *category*. Error *severity* (either *major* or *minor*) is assigned independently of category. Spans with no marked errors have *neutral* severity and no category. Possible error categories are displayed in Table 7.

You will be assessing translations at the segment level, where a segment may contain one or more sentences. Each segment is aligned with a corresponding source segment, and both segments are displayed within their respective documents. Annotate segments in natural order, as if you were reading the document. You may return to revise previous segments.

Please identify all errors within each translated segment, up to a maximum of five. If there are more than five errors, identify only the five most severe. If it is not possible to reliably identify distinct errors because the translation is too badly garbled or is unrelated to the source, then mark a single *Non-translation* error that spans the entire segment.

To identify an error, highlight the relevant span of text, and select a category/sub-category and severity level from the available options. (The span of text may be in the source segment if the error is a source error or an omission.) When identifying errors, please be as fine-grained as possible. For example, if a sentence contains two words that are each mistranslated, two separate mistranslation errors should be recorded. If a single stretch of text contains multiple errors, you only need to indicate the one that is most severe. If all have the same severity, choose the first matching category listed in the error typology (eg, *Accuracy*, then *Fluency*, then *Terminology*, etc).

Please pay particular attention to document context when annotating. If a translation might be questionable on its own but is fine in the context of the document, it should not be considered erroneous; conversely, if a translation might be acceptable in some context, but not within the current document, it should be marked as wrong.

There are two special error categories: *Source error* and *Non-translation*. Source errors should be annotated separately, highlighting the relevant span in the source segment. They do not count against the five-error limit for target errors, which should be handled in the usual way, whether or not they resulted from a source error. There can be at most one *Non-translation* error per segment, and it should span the entire segment. No other errors should be identified if *Non-Translation* is selected.

Table 6: MQM annotator guidelines

Since MQM doesn't ask annotators for quality scores, those scores are derived automatically from the identified error spans and their classifications, based on a *weighting* of each error severity and category. Table 8 summarizes this weighting scheme, in which segment-level scores can range from 0 (perfect) to 25 (worst). The final segment-level score is an average over scores from all annotators. In some settings (e.g. calculating correlation for learned metrics), the scores are negated.

We use the same weighting to obtain scores from errors identified by AUTOMQM.

B Datasets' Statistics

See Table 9 for a summary of the number of systems and annotated segments per system in the evaluation datasets used in this work.

C Score Prediction Prompt

Figure 9 contains the GEMBA-SQM prompt that we used for our 0-shot experiments.

D Sampling in-context learning examples for AutoMQM

Figure 10 shows the rejection criteria used when sampling example sets as discussed in §4.

E Additional Results

Figures 11, 12, 13 and 8 present additional experimental results.

Error Category		Description
Accuracy	Addition	Translation includes information not present in the source.
	Omission	Translation is missing content from the source.
	Mistranslation	Translation does not accurately represent the source.
	Untranslated text	Source text has been left untranslated.
Fluency	Punctuation	Incorrect punctuation (for locale or style).
	Spelling	Incorrect spelling or capitalization.
	Grammar	Problems with grammar, other than orthography.
	Register	Wrong grammatical register (eg, inappropriately informal pronouns).
	Inconsistency	Internal inconsistency (not related to terminology).
	Character encoding	Characters are garbled due to incorrect encoding.
Terminology	Inappropriate for context	Terminology is non-standard or does not fit context.
	Inconsistent use	Terminology is used inconsistently.
Style	Awkward	Translation has stylistic problems.
Locale convention	Address format	Wrong format for addresses.
	Currency format	Wrong format for currency.
	Date format	Wrong format for dates.
	Name format	Wrong format for names.
	Telephone format	Wrong format for telephone numbers.
	Time format	Wrong format for time expressions.
Other		Any other issues.
Source error		An error in the source.
Non-translation		Impossible to reliably characterize distinct errors.

Table 7: MQM hierarchy.

```
Score the following translation from {src_lang} to {tgt_lang} with respect to the
human reference on a continuous scale from 0 to 100 that starts with "No meaning
preserved", goes through "Some meaning preserved", then "Most meaning preserved
and few grammar mistakes", up to "Perfect meaning and grammar".

{src_lang} source: "{source}"
{tgt_lang} human reference: "{reference}"
{tgt_lang} translation: "{candidate}"
Score (0-100): {score}
```

Figure 9: The *score prediction* prompt used in this paper. Equivalent to the `GEMBA-SQM` prompt in Kocmi and Federmann (2023). Parts in purple are only included for *reference-based* evaluation, while parts in orange represent slots for outputs and are only included for in-context examples.

Severity	Category	Weight
Major	Non-translation	25
	all others	5
Minor	Fluency/Punctuation	0.1
	all others	1
Neutral	all	0

Table 8: MQM error weighting.

LP	#Sys	#Seg
en→de	13	1315
zh→en	14	1875
en→ru	15	1315

LP	#Sys	#Seg
en→kk	11	998
kk→en	11	1000
en→gu	11	998
gu→en	11	1016

Table 9: The number of systems and segments that have MQM scores (left) and DA scores (right) used as ground-truth in this work.

```python
def check_icl_set(
    examples: pd.DataFrame,
    min_errors=3,
    majmin_threshold=2,
    cat_diversity=2,
    min_clen=20,
    max_clen=400,
):
    # Check if they have the same number of spans as severity/category
    if not examples.apply(
        lambda r:
          len(r['span']) == len(r['severity']) and len(r['span']) == len(r['category']),
        axis=1
    ).all():
      return False

    # Check if there are at least min_errors
    if examples['severity'].apply(lambda svs: len(svs)).sum() < min_errors:
      return False

    # Check that there's a balance of major and minor errors.
    major_count = examples['severity'].apply(lambda svs: sum([s=='major' for s in svs])).sum()
    minor_count = examples['severity'].apply(lambda svs: sum([s=='minor' for s in svs])).sum()
    if abs(major_count - minor_count) > majmin_threshold:
        return False

    # Check that at least cat_diversity error types are represented.
    categories = examples['category'].apply(lambda cs: [c.split("/")[0] for c in cs])
    represented_error_types = set().union(*categories.tolist())
    if len(represented_error_types) < cat_diversity:
        return False

    top_clen = examples.apply(
        lambda row: max(len(row[s]) for s in ('source', 'reference', 'candidate')
    ), axis=1).max()
    bot_clen = examples.apply(
        lambda row: min(len(row[s]) for s in ('source', 'reference', 'candidate')),
    axis=1).min()

    if top_clen > max_clen or bot_clen < min_clen:
      return False

    # All checks passed.
    return True
```

Figure 10: Rejection criteria used when sampling *in-context learning* examples for AUTOMQM.

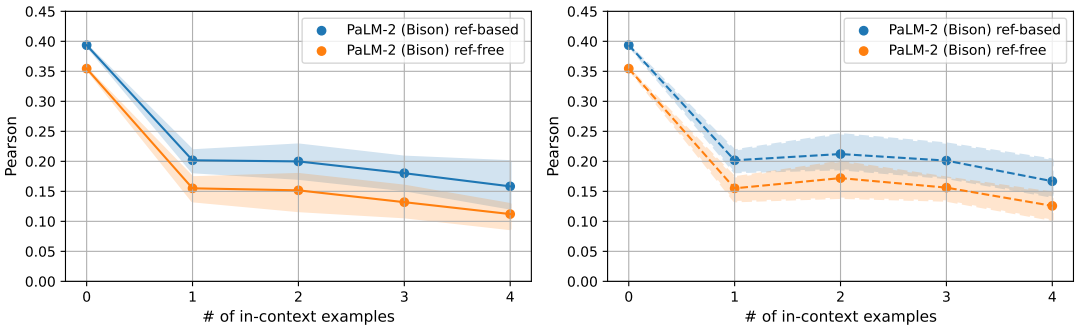


Figure 11: Mean *Pearson* and its interquartile range (IQR), as we increase the number of in-context examples in the *score prediction* prompt, sampled with *uniform* (left) and *stratified* (right) sampling, for WMT22 EN-DE.

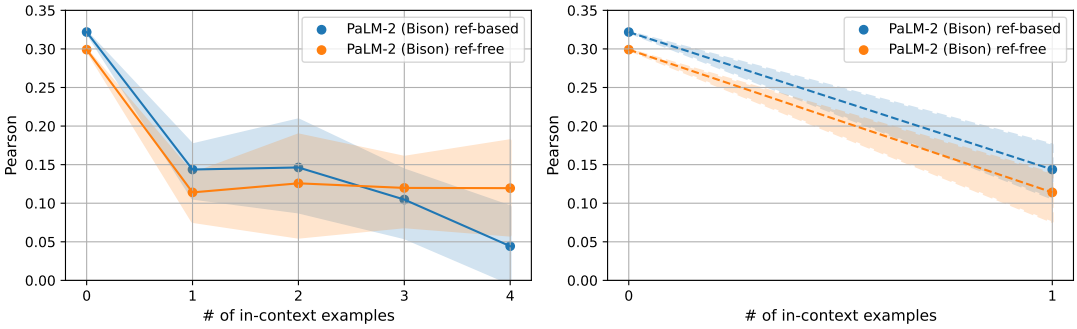


Figure 12: Mean *Pearson* and its interquartile range (IQR), as we increase the number of in-context examples in the *score prediction* prompt, sampled with *uniform* (left) and *stratified* (right) sampling, for WMT22 ZH-EN.

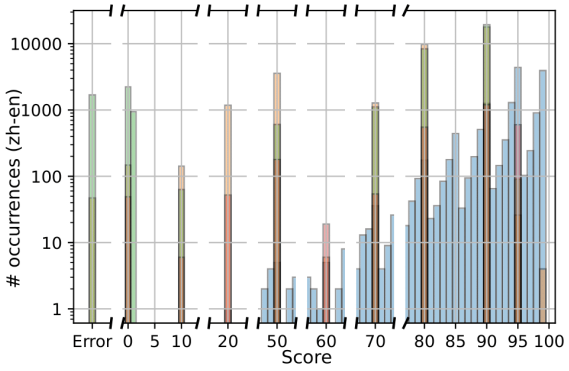


Figure 13: Distribution of scores for various LLM *reference-based* evaluators, on the ZH-EN test set. Note that the y axis is in *log-scale*.

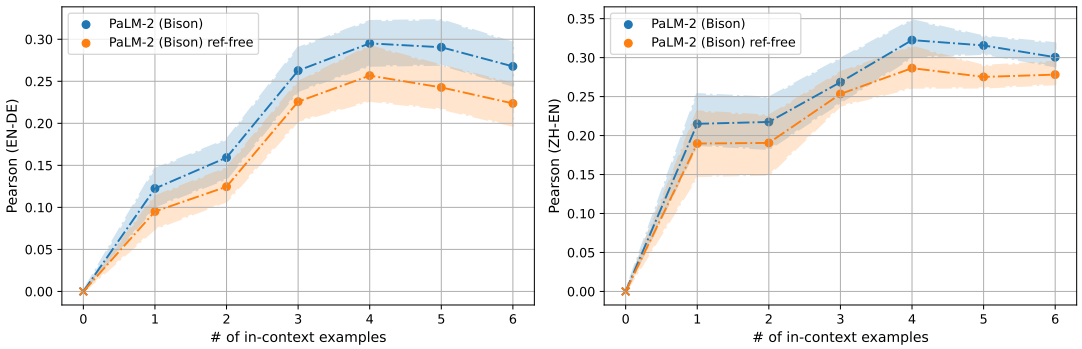


Figure 14: Mean *Pearson* and its interquartile range (IQR), as we increase the number of in-context examples in the AUTOMQM prompt, for EN-DE (left) and ZH-EN (right).

Author Index