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. ©2023 Association for Computational Linguistics

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 bidirection 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

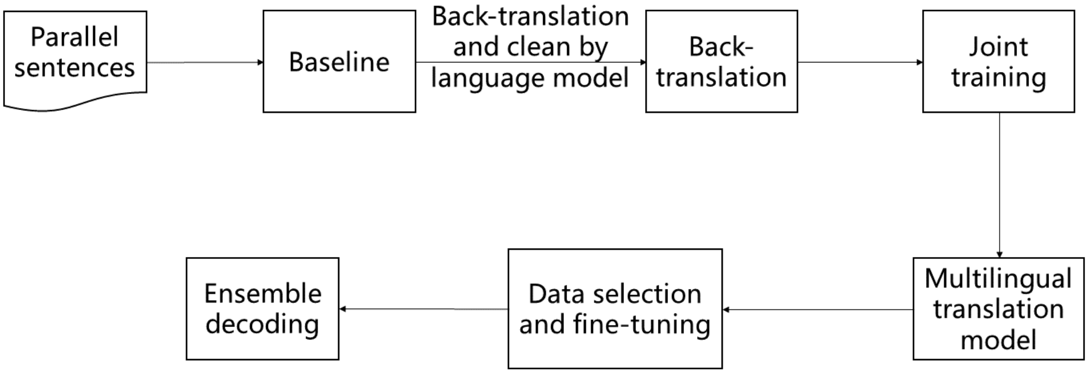

Figure 1: The work flow of GTCOM machine translation competition systems

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.