Findings of the 2021 Conference on Machine Translation (WMT21)

Farhad Akhbardeh RIT
Arkady Arkhangorodsky DiDi Labs
Magdalena Biesialska UPC
Ondřej Bojar Charles University
Rajen Chatterjee Apple Inc.
Vishrav Chaudhary Facebook AI
Marta R. Costa-jussà UPC
Cristina España-Bonet DFKI
Angela Fan Facebook AI
Christian Federmann Microsoft Cloud + AI
Markus Freitag Google Research
Yvette Graham Trinity College Dublin
Roman Grundkiewicz Microsoft
Barry Haddow University of Edinburgh
Leonie Harter DFKI
Kenneth Heafield University of Edinburgh
Christopher M. Homan RIT
Matthias Huck SAP SE
Kwabena Amponsah-Kaakyire DFKI
Jungo Kasai U. of Washington
Daniel Khashabi Allen Institute for AI
Kevin Knight DiDi Labs
Tom Kocmi Microsoft
Philipp Koehn JHU
Nicholas Lourie New York University
Christof Monz University of Amsterdam
Makoto Morishita NTT
Masaaki Nagata NTT
Ajay Nagesh DiDi Labs
Toshiaki Nakazawa University of Tokyo
Matteo Negri FBK
Santanu Pal WIPRO AI
Allahsera Tapo RIT
Marco Turchi FBK
Valentin Vydrin INALCO
Marcos Zampieri RIT

Abstract

This paper presents the results of the news translation task, the multilingual low-resource translation for Indo-European languages, the triangular translation task, and the automatic post-editing task organised as part of the Conference on Machine Translation (WMT) 2021. In the news task, participants were asked to build machine translation systems for any of 10 language pairs, to be evaluated on test sets consisting mainly of news stories. The task was also opened up to additional test suites to probe specific aspects of translation. In the Similar Language Translation (SLT) task, participants were asked to develop systems to translate between pairs of similar languages from the Dravidian and Romance family as well as French to two similar low-resource Manding languages (Bambara and Maninka). In the Triangular MT translation task, participants were asked to build a Russian to Chinese translator, given parallel data in Russian-Chinese, Russian-English and English-Chinese. In the multilingual low-resource translation for Indo-European languages task, participants built multilingual systems to translate among Romance and North-Germanic languages. The task was designed to deal with the translation of documents in the cultural heritage domain for relatively low-resourced languages. In the automatic post-editing (APE) task, participants were asked to develop systems capable to correct the errors made by an unknown machine translation systems.

1 Introduction

The Sixth Conference on Machine Translation (WMT21)[1] was held online with EMNLP 2021 and hosted a number of shared tasks on various aspects of machine translation. This conference built on 15 previous editions of WMT as workshops and conferences (Koehn and Monz, 2006; Callison-Burch et al., 2007, 2008, 2009, 2010, 2011, 2012; Bojar et al., 2013, 2014, 2015, 2016, 2017, 2018a; Barrault et al., 2019, 2020).

This year we conducted several official tasks. In this paper we report on the news task, the multilingual low-resource translation for Indo-European languages task, the triangular translation task, and the automatic post-editing task. Additional shared tasks are described in separate papers in these proceedings:

[1] http://www.statmt.org/wmt21/

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 1–88
November 10–11, 2021.

- biomedical translation (Yeganova et al., 2021)
- efficiency (Heafield et al., 2021)
- large-scale multilingual machine translation (Wenzek et al., 2021)
- machine translation using terminologies (Alam et al., 2021)
- metrics (Freitag et al., 2021b)
- quality estimation (Specia et al., 2021)
- unsupervised and very low-resource translation (Libovický and Fraser, 2021)

In the news translation task (Section 2), participants were asked to translate a shared test set, optionally restricting themselves to the provided training data (“constrained” condition). We included 20 translation directions this year, with translation between English and each of Chinese, Czech, German, Japanese and Russian, as well as French↔German being repeated from last year, and English to and from Hausa and Icelandic being new for this year, along with Bengali↔Hindi and Xhosa↔Zulu. The translation tasks covered a range of language families, and included both low-resource and high-resource pairs. System outputs for each task were evaluated both automatically and manually, but we only include the manual evaluation here.

The human evaluation (Section 3) involves asking human judges to score sentences output by anonymized systems. We obtained large numbers of assessments from researchers who contributed evaluations proportional to the number of tasks they entered. We collected additional assessments from a pool of linguists, as well as crowd-workers. This year, the official manual evaluation metric is again based on judgments of adequacy on a 100-point scale, a method (known as “direct assessment”, DA) that we explored in the previous years with convincing results in terms of the trade-off between annotation effort and reliable distinctions between systems. In addition, other golden standards with this year’s systems were collected. The human-in-the-loop GENIE leaderboard (Khashabi et al., 2021) conducted de→en evaluations independently in a Likert scale (Section 3.5). We refer the reader to Freitag et al. (2021b) for MQM scoring of en→de, en→ru, and zh→en.

The primary objectives of WMT are to evaluate the state of the art in machine translation, to disseminate common test sets and public training data with published performance numbers, and to refine evaluation and estimation methodologies for machine translation. As before, all of the data, translations, and collected human judgments are publicly available.[2] We hope these datasets serve as a valuable resource for research into data-driven machine translation, automatic evaluation, or prediction of translation quality. News translations are also available for interactive visualization and comparison of differences between systems at `http://wmt.ufal.cz/` using MT-ComparEval (Sudarikov et al., 2016), and also on ExplainaBoard[3] (Liu et al., 2021b).

In order to gain further insight into the performance of individual MT systems, we again organized a call for dedicated “test suites”. Test suites are custom additions to the inputs. Anyone can provide a test suite for any subset of news translation task languages and we ensure that the test suite is requested from all participating MT systems. The MT outputs are delivered back to test suite authors for evaluation, which can be manual, automatic or both, focusing on any possible aspect of the MT systems. This year, five test suites were acquired and translated by participating MT systems but only two were then analyzed in time for these proceedings:

- Freitag et al. (2021b), the metrics task paper, used TED talks as additional domain, scored them with MQM, and further used these outputs and scores to assess domain-dependence of MT evaluation metrics.

- Macketanz et al. (2021) reports on the fourth application of a fine-grained test suite for German↔English linguistic phenomena. The previous instances (Macketanz et al., 2018; Avramidis et al., 2019, 2020) use the same underlying collection of sentences and thus allow to observe the overall development of MT systems in clear categories. This year, the major jump was observed in the category of idioms, especially due to a few exceptional MT systems. Many phenomena are being solved almost perfectly, the difficult categories remain false friends, ambiguity and multi-word expressions.

The goal of the Similar Language Translation (SLT) task (Section 4) is to evaluate the perfor-

[2] `http://statmt.org/wmt21/results.html`
[3] `http://explainaboard.nlpedia.ai/leaderboard/task-mt/index.php`

mance of MT systems taking into account the similarity between pairs of closely-related languages from the same language family. Following the interest of the community in this topic (Costa-jussà et al., 2018; Popović et al., 2020) and the success of the past two editions of the SLT task task at WMT 2019 and WMT 2020, we organize a third iteration of the task at WMT 2021. SLT 2021 features a pair of similar Dravidian languages, namely Tamil - Telugu, and multiple pairs of Romance languages involving Catalan, Spanish, Portuguese, and Romanian in all possible combinations. A new track with French and two similar low-resource Manding languages: Bambara and Maninka was also included to encourage participants to take advantage of the similarity between Bambara and Maninka and explore data augmentation techniques, a typical scenario of low-resource languages. Finally, translations were evaluated in both directions using three automatic metrics: BLEU, RIBES, and TER.

The primary goals of the Triangular MT task (Section 5) are to promote translation between non-English languages, to optimally mix direct and indirect parallel resources and exploit noisy web data sources to build an MT system. Specifically, the task was Russian to Chinese machine translation, given parallel data comprising of direct (Russian-Chinese) and indirect (Russian-English and English-Chinese) sources. The submitted systems were evaluated on a (secret) mixed-genre test set, drawn from the web and curated manually for high-quality segment pairs.

The multilingual low-resource translation for Indo-European languages task (MLLR, Section 6) aims to investigate the best approaches to deal with multilingual translation. Usually, multilingual translation is done with the help of a high-resourced language, e.g. English. In MLLR, we evaluate translation quality for Icelandic–Norwegian Bokmål–Swedish (North-Germanic) and Catalan–Italian–Occitan–Romanian (Romance). Higher resourced languages (Danish, German, English, Spanish, French and Portuguese) are allowed for training but not evaluated. We focus on a specific domain: cultural heritage documents are extracted from Europeana and Wikipedia, a domain where named entities may also play a role in translation quality. The evaluation is done at language family level with a combination of automatic metrics (BLEU, TER, chrF, BertScore and COMET) and complemented by a manual evaluation on a subset of language pairs.

The automatic post-editing (APE) task (Section 7) focuses on another MT-related problem: the correction of machine-translated text generated by an unknown system. In continuity with last year, in this seventh iteration of the task at WMT we focused on two language pairs (English-German and English-Chinese), using data drawn from English Wikipedia articles and translated with neural MT systems. The evaluation was carried out both automatically – with TER and BLEU respectively used as primary and secondary metric - and manually – with the same direct assessment method used for the news translation task.

2 News Translation Task

This recurring WMT task assesses the quality of MT on text from the news domain. As in the previous year, we included Chinese, Czech, German, Japanese and Russian (to and from English) as well as French↔German. New language pairs for this year were Icelandic and Hausa (to and from English) as well as Bengali↔Hindi and Xhosa↔Zulu.

2.1 Test Data

As in previous years, the test sets consist of unseen translations prepared specially for the task. The test sets are publicly released to be used as translation benchmarks in the coming years. Here we describe the production and composition of the test sets.

The source texts for the test sets were all extracted from online news sites, with the exception of Bengali↔Hindi and Xhosa↔Zulu, which were part of the FLORES-101 benchmark (Goyal et al., 2021) and extracted from Wikipedia. The sources used for the online news are shown in Table 1, and all articles are from the second half of 2020. For the French↔German task, we specifically selected financial and economic news, whereas for the other news sources, we randomly selected articles from general online news, including politics, sports, international and local events.

For all language pairs, we aimed for a test set size of 1000 sentences, and to ensure that the test sets were "source-original", in that the source text is the original article and the target text is the translation. This is to avoid "translationese" effects on the source language, which can have a detrimental effect on the accuracy of evaluation (Freitag et al., 2019; Laubli et al., 2020; Graham et al., 2020). The exceptions were Chinese→English, where we used a larger test set of 1948 sentences, and the FLORES-101 test sets which were around 500 sentences, and derived from English source documents. For language pairs that were new this year (i.e. Icelandic↔English and Hausa↔English) we prepared development sets using the same process as the test set, but concatenating both translation directions into the same set. For each translated article in the development set, the direction of translation is clearly identified.

For WMT20, we experimented with using test sources with line (segment) boundaries at paragraphs (not sentences) for some language pairs, but we found no evidence that translators used their new freedom to reorganise sentences, and the longer lines possibly made evaluation more difficult, so we reverted to a sentence-per-line format this year. For selected language sources (Czech, German and English, when translated into the recurring languages) we retained the paragraph boundaries from the original articles, but within the paragraphs, the sentences were in separate segments. It was up to the participating systems to make use of the paragraph breaks or not, but the systems were expected to preserve the segment boundaries.

The test sets for WMT21 were released using a new XML format, replacing the "pseudo xml" SGML format which had been used for many years. The advantages of the new format are: (i) it can be processed with standard XML tools, and there is no longer any doubt about how to treat special XML characters such as the ampersand ("&"); (ii) the source, all references and all submissions can be contained in one convenient XML file; (iii) the metadata better matches the needs of the task, and can be extended as necessary. We created simple tools for converting from text-based files to the new XML format.[4]

The translation of the test sets was performed by professional translation agencies, according to the brief supplied in Appendix B. Several language pairs got special attention. For Chinese↔English, Russian↔English and German↔English, we obtained a second reference in each direction from a different translation agency, labelled "B". For German↔English, the "B" reference was found to be a post-edited version of one of the participating online systems, so we had to discard it. Microsoft then sponsored a third independent translation, labelled "C", and the metrics task organizers with the support from Google later provided yet another German↔English reference, discussed only in Freitag et al. (2021b) as "D". For Czech↔English, the first reference (labelled "A") which served in reference-based manual evaluations, was provided by a translation agency in both directions. The second Czech↔English reference (labelled "B") which served as another system in the competition was provided by professional translators recruited from teachers and students of translation studies into Czech and three students and graduates of translation studies and one translator, English native speaker, into English.

2.2 Training Data

As in past years we provided a selection of parallel and monolingual corpora for model training, and development sets to tune system parameters. Participants were permitted to use any of the provided corpora to train systems for any of the language pairs. As well as providing updates on many of the previously released data sets, we included several new data sets, mainly to support the new language pairs.

Our training data includes the latest version of ParaCrawl (Bañón et al., 2020) for all language pairs where it is available. New for this year is a ParaCrawl corpus for Chinese↔English, which contains 14M sentences, as well as a small Hausa↔English ParaCrawl. The JParaCrawl corpus (for Japanese↔English) is constructed in a similar way to ParaCrawl, but by a different group (Morishita et al., 2020).

For Icelandic↔English we used the recently released ParIce (Barkarson and Steingrímsson, 2019) a source of parallel data, and the Icelandic Gigaword corpus for monolingual data (Steingrímsson et al., 2018).

For Hausa↔English, the data was mainly drawn from Opus (Tiedemann and Nygaard, 2004), which is mostly religious and IT localisation text. We added a small ($<$ 6000) parallel sentence corpus extracted from the website of Ayatollah Khamenei,[5] now only accessible using the

[4] https://github.com/wmt-conference/wmt-format-tools

[5] https://english.khamenei.ir/

English	ABC News (5), Al Jazeera (1), All Africa (2), BBC (4), Brisbane Times (3), CBS LA (1), CBS News (3), CNBC (1), CNN (1), Daily Express (4), Daily Mail (1), Egypt Independent (3), Fox News (2), Guardian (6), LA Times (1), London Evening Standard (2), Metro (1), NDTV (7), New York Times (2), RTE (1), Russia Today (5), Seattle Times (4), Sky (1), The Independent (1), The Sun (2), UPI (1), VOA (1), news.com.au (1), novinite.com (1),
Chinese	China News (76), Hunan Ribao (5), Jingji Guancha Bao (3), Macao Government (2), Nhan Dan (3), RFI Chinese (6), VOA Chinese (3), Xinhua (57), tsrus.cn (1),
Czech	Aktuálně (4), Blesk (5), Denik (3), Dnes (1), E15 (1), Haló noviny (5), Hospodářské Noviny (1), Idnes (2), Lidovky (7), Mediafax (6), Novinky (6), Týden (1), Tydenek Homer Mostecka (1), ČT24 (4), Česká Pozice (6), Česká Televize (4), České Noviny (4), Český Rozhlas (1),
German	Aachener Nachrichten (1), Abendzeitung Mn̈chen (1), Abendzeitung Nürnberg (1), Allgemeine Zeitung (1), Augsburger-allgemeine (1), Braunschweiger Zeitung (1), Das Bild (3), Dresdner Neueste Nachrichten (1), Euronews (1), Frankfurter Allgemeine Zeitung (1), Freie Presse (1), Handelsblatt (1), Hessische/Niedersaechsische Allgemeine (1), Infranken (3), Kurier (2), Lampertheimer Zeitung (3), Landeszeitung (1), Main-Netz (1), Mainpost (1), Mittelbayerische Zeitung (2), Mitteldeutsche Zeitung (2), Morgenpost (2), Neue Presse (Coburg) (2), Nordbayerischer Kurier (3), OE24 (1), Passauer Neue Presse (2), Peiner Allgemeine Zeitung (2), Pforzheimer Zeitung (1), Potsdamer Neueste Nachrichten< (1), Rhein Zeitung (2), Rundschau online (1), Söster Anzeiger (1), Salzburger Nachrichten (1), Schwäbische (2), Schwäbische post (2), Schwarzwälder Bote (2), Tiroler Tageszeitung (2), Usinger Anzeiger (1), Westfälische Nachrichten (2), Wienerzeitung (1),
Hausa	Deutsche Welle (7), Freedom radio (22), Leadership (19), Premium Times (20), RFI Hausa (10), VOA Hausa (18), VON Hausa (4),
Japanese	Fukui Shimbun (1), Hokkaido Shimbun (5), Iwate Nippo (3), Saga Shimbun (3), Sanyo Shimbun (4), Shizuoka Shimbun (11), Ube nippo Shimbun (2), Yaeyama mainichi shimbun (1), Yahoo (49), Yamagata Shimbun (2),
Russian	Altapress (1), Altyn-orda (1), Argumenti Nedely (5), Argumenty i Fakty (6), Armenpress (1), BBC Russian (1), Delovoj Peterburg (1), ERR (5), Gazeta (4), Interfax (3), Izvestiya (11), Kommersant (1), Komsomolskaya Pravda (7), Lenta (6), Lgng (2), Moskovskij Komsomolets (9), Novye Izvestiya (1), Ogirk (1), Parlamentskaya Gazeta (3), Rossiskaya Gazeta (5), Russia Today (8), Russkaya Planeta (1), Sovsport (2), Sport Express (9), Tyumenskaya Oblast Segodnya (1), VOA Russian (1), Vedomosti (2), Vesti (6), Xinhua (3),
German (economic)	Aachener Nachrichten (1), Abendzeitung Mn̈chen (1), Das Bild (1), Der Spiegel (2), Epoch Times (1), Frankfurter Allgemeine Zeitung (6), Handelsblatt (17), Haz (2), Kurier (4), Lübecker Nachrichten (1), Mindener Tageblatt (1), Mittelbayerische Zeitung (1), NZZ (1), Neue Westfälische (1), Onetz (1), Passauer Neue Presse (2), Rheinische Post (1), Russia Today (3), Süddeutsche Zeitung (8), Salzburger Nachrichten (2), Tiroler Tageszeitung (1), Volksstimme (1), Yahoo (1), come-on.de (1),
French (econmic)	Algérie Presse Service (3), Aujourd'hui le Maroc (5), Dernière Heure (4), Dernières Nouvelles d'Alsace (1), Euronews (2), L'Independant (1), L'express (2), La Croix (4), La Meuse (3), La Tribune (4), La Venir (1), Le Devoir (3), Le Figaro (17), Le Monde (5), Le Quotidien (1), Les Echos (1), Liberté Algerie (1), Libre Belgium (1), Madagascar tribune (1), Metro Canada (1), Nice Matin (1), Nouvel Obs (6), Russia Today (4), VOA Afrique (2),

Table 1: Composition of the test sets. The economic arcticles were used for French↔German only. We did not record the sources for the Icelandic articles, and the Bengali, Hindi, Xhosa and Zulu articles were from Wikipedia.

Europarl Parallel Corpus

	Czech ↔ English		German ↔ English		German ↔ French	
Sentences	645,241		1,825,745		1,801,076	
Words	14,948,900	17,380,340	48,125,573	50,506,059	47,517,102	55,366,136
Distinct words	172,452	63,289	371,748	113,960	368,585	134,762

News Commentary Parallel Corpus

	Czech ↔ English		German ↔ English		Russian ↔ English	
Sentences	253,456		388,813		331,596	
Words	5,674,011	6,270,051	9,921,515	9,840,910	8,469,701	8,820,805
Distinct words	176,403	70,774	215,101	86,518	207,701	82,938

	Chinese ↔ English		Japanese ↔ English		German ↔ French	
Sentences	313,934		1,851		296,022	
Words	–	7,982,550	–	45,438	7,671,513	9,346,818
Distinct words	–	76,372	–	6,280	185,348	87,481

Common Crawl Parallel Corpus

	German ↔ English		Czech ↔ English		Russian ↔ English		French ↔ German	
Sentences	2,399,123		161,838		878,386		622,288	
Words	54,575,405	58,870,638	3,529,783	3,927,378	21,018,793	21,535,122	13,991,973	12,217,457
Distinct words	1,640,835	823,480	210,170	128,212	764,203	432,062	676,725	932,137

ParaCrawl Parallel Corpus

	German ↔ English		Czech ↔ English		Chinese ↔ English	
Sentences	82,638,202		14,083,311		14,170,585	
Words	1,543,410,882	1,613,780,145	240,233,151	260,801,934	–	253,776,811
Distinct Words	15,256,769	7,765,311	2,655,118	1,972,030	–	1,871,639

	Japanese ↔ English		Russian ↔ English		French ↔ German	
Sentences	10,120,013		12,654,509		7,222,574	
Words	–	274,368,443	232,950,488	266,368,340	145,190,707	123,205,701
Distinct Words	–	2,051,246	2,913,181	1,816,590	1,534,068	2,368,682

	Icelandic ↔ English		Hausa ↔ English	
Sentences	2,392,422		158,968	
Words	39,528,080	42,454,372	4,041,027	3,957,605
Distinct Words	709,945	416,986	102,962	101,049

EU Press Release Parallel Corpus

	Czech ↔ English		German ↔ English	
Sentences	452,411		1,631,639	
Words	7,214,324	7,748,940	26,321,432	27,018,196
Distinct words	141,077	83,733	402,533	197,030

Yandex 1M Parallel Corpus

	Russian ↔ English	
Sentences	1,000,000	
Words	24,121,459	26,107,293
Distinct	701,809	387,646

CzEng v2.0 Parallel Corpus

	Czech ↔ English	
Sentences	60,980,645	
Words	757,316,261	848,016,692
Distinct	3,684,081	2,493,804

WikiTitles Parallel Corpus

	Chinese ↔ English		Czech ↔ English		German ↔ English		Hausa ↔ English	
Sentences	922,194		410,977		1,474,196		7,501	
Words	–	2,549,611	990,191	1,065,417	3,219,123	3,763,461	14,285	14,629
Distinct	–	380,234	218,992	186,375	674,927	573,280	7,855	7,827

	Icelandic ↔ English		Japanese ↔ English		Russian ↔ English		German ↔ French	
Sentences	50,181		757,052		1,189,097		1,006,563	
Words	90,620	100,847	–	2,016,400	3,244,102	3,261,299	2,142,193	2,543,265
Distinct	40,570	34,440		281,880	534,392	457,933	503,342	444,330

Figure 1: Statistics for the training sets used in the translation task. The number of words and the number of distinct words (case-insensitive) is based on the Moses tokenizer and IndicNLP (https://github.com/anoopkunchukuttan/indic_nlp_library).

CCMT Corpus

	casia2015	casict2011	casict2015	datum2011	datum2017	neu2017
Sentences	1,050,000	1,936,633	2,036,834	1,000,004	999,985	2,000,000
Words (en)	20,571,578	34,866,598	22,802,353	24,632,984	25,182,185	29,696,442
Distinct words (en)	470,452	627,630	435,010	316,277	312,164	624,420

Extra Japanese-English Parallel Data

	Subtitles		Kyoto		TED	
Sentences	2,801,388		443,849		223,108	
Words	–	23,933,060	–	11,622,252		4,554,409
Distinct	–	161,484	–	191,885	–	60,786

Extra Hausa-English Parallel Data

	Khamenei		Opus	
Sentences	5,837		584,004	
Words	217,543	167,466	8,385,179	8,994,622
Distinct	6,075	7,942	219,203	193,518

CC-Aligned

	Bengali↔Hindi		Xhosa↔Zulu	
Sentences	3,365,142		94,323	
Words	40,782,432	45,609,689	1,689,086	1,658,266
Distinct	996,612	860,033	186,070	173,148

United Nations Parallel Corpus

	Russian ↔ English		Chinese ↔ English	
Sentences	23,239,280		15,886,041	
Words	570,099,284	601,123,628	–	425,637,920
Distinct	1,446,782	1,027,143	–	769,760

Synthetic parallel data (both directions combined)

	Czech ↔ English		Russian ↔ English		Chinese ↔ English	
Sentences	126,828,081		76,133,209		19,763,867	
Words	2,351,230,606	2,655,779,234	1,511,996,711	1,698,428,744	–	416,567,173
Distinct	5,745,323	3,840,231	5,928,141	3,889,049	–	1,188,933

Wikimatrix Parallel Data

	Czech ↔ English		German ↔ English		Japanese ↔ English		Icelandic ↔ English	
Sentences	2,094,650		6,227,188		3,895,992		313,875	
Words	34,801,119	39,197,172	113,445,806	118,077,685	–	72,320,248	5,395,042	6,475,011
Distinct	1,068,844	798,095	2,855,263	1,827,785	–	1,106,529	328,369	231,192

	Russian ↔ English		Chinese ↔ English		German ↔ French	
Sentences	5,203,872		2,595,119		3,350,816	
Words	93,828,313	102,937,537	–	58,615,891	68,249,384	59,422,699
Distinct	2,233,043	1,592,190	–	1,059,537	1,067,450	1,844,533

Figure 2: Statistics for the training sets used in the translation task. The number of words and the number of distinct words (case-insensitive) is based on the Moses tokenizer and IndicNLP (https://github.com/anoopkunchukuttan/indic_nlp_library).

News Language Model Data

	English	German	Czech	Russian	Japanese
Sentences	274,929,980	386,987,716	97,396,609	111,118,861	14,389,733
Words	6,782,988,670	7,951,191,279	1,760,715,133	2,010,171,968	–
Distinct words	8,329,647	39,524,377	5,960,637	5,679,507	–

	Icelandic	Chinese	French	Hausa	Hindi	Bengali
Sentences	534,647	10,771,382	96,402,399	272,966	46,187,245	10,101,626
Words	9,653,929	–	2,338,364,059	7,305,501	872,106,937	148,586,981
Distinct words	308,924	–	3,975,116	125,350	2,752,071	1,091,788

Document-Split News LM Data (not dedupped)

	Czech	English	German
Sentences	142,478,129	531,904,913	739,041,709
Words	2,221,995,079	11,472,609,712	12,524,314,673
Distinct words	5,744,574	8,595,778	26,849,693

Common Crawl Language Model Data

	English	German	Czech	Russian
Sent.	3,074,921,453	2,872,785,485	333,498,145	1,168,529,851
Words	65,104,585,881	65,147,123,742	6,702,445,552	23,332,529,629
Dist.	342,149,665	338,410,238	48,788,665	90,497,177

	Chinese	Icelandic	Hausa	French
Sent.	1,672,324,647	24,627,579	1,467,326	4,898,012,445
Words	–	595,998,326	20,082,665	126,364,574,036
Dist.	–	7,483,421	688,610	363,878,959

Figure 3: Statistics for the monolingual training sets used in the translation task. The number of words and the number of distinct words (case-insensitive) is based on the Moses tokenizer and IndicNLP (https://github.com/anoopkunchukuttan/indic_nlp_library).

Test Sets

	Czech → EN			EN → Czech			German → EN			EN → German		
Lines.	1000			1002			1000			1002		
Words	17,914	22,080	22,570	27,454	25,907	27,190	18,190	20,668	20,541	27,454	28,273	28,673
Distinct words	6,457	4,032	4,425	5,374	8,295	8,577	5,115	4,012	3,980	5,374	6,841	6,697

	Chinese → EN		EN → Chinese			Russian → EN			EN → Russian		
Lines.	1948		1002			1000			1002		
Words	–	72,334	27,454	–	–	17,796	21,400	21,185	27,454	26,413	26,253
Distinct words	–	8,290	5.374	–	–	6,315	4,214	4,230	5,374	8,591	8,377

	Icelandic → EN		EN → Icelandic		Japanese → EN		EN → Japanese		Hausa ↔ EN	
Lines.	1000		1000		1005		1000		997	
Words	19,930	22,749	26,467	25,557	–	28,846	26,467	–	31,362	27,519
Distinct words	5,282	3,773	5,258	6,614	–	5,001	5,258	–	4,032	4,240

	EN ↔ Hausa		Bengali → Hindi		Hindi → Bengali		Xhosa → Zulu		Zulu ↔ Xhosa	
Lines.	1000		503		509		503		509	
Words	26,467	33,915	11,439	14,133	14,286	11,136	9,180	9,314	9,320	9,065
Distinct words	5,258	4,713	4,514	3,686	3,402	4,091	5,499	5,265	4,961	5,093

	French → German		German → French	
Lines.	1026		1000	
Words	30,143	26,353	18,801	26,407
Distinct words	5,395	6,021	5,198	4,613

Figure 4: Statistics for the test sets used in the translation task. In the cases that there are three word counts, these are for source, first target translation, and second target translation. The number of words and the number of distinct words (case-insensitive) is based on the Moses tokenizer and IndicNLP (https://github.com/anoopkunchukuttan/indic_nlp_library).

Wayback Machine.[6]

For the two FLORES-101 language pairs (i.e. Bengali↔Hindi and Xhosa↔Zulu) all training data is from the CC-Aligned corpus (El-Kishky et al., 2020).

Other language pairs used the same data sets as last year, with updates wherever available.

The monolingual data we provided was similar to last year's, with a 2020 news crawl[7] added to all the news corpora. Note that news crawl now includes 59 languages, so is not limited to languages used in WMT. In addition, we provided versions of the news corpora for Czech, English and German, with both the document and paragraph structure retained. In other words, we did not apply sentence splitting to these corpora, and we retained the document boundaries and text ordering of the originals.

Some statistics about the training and test materials are given in Figures 1, 2, 3 and 4.

2.3 Submitted Systems

In 2021, we received a total of 173 submissions. The participating institutions are listed in Table 2 and detailed in the rest of this section. Each system did not necessarily appear in all translation tasks. We also included online MT systems (originating from 5 services), which we anonymized as ONLINE-A,B,G,W,Y. All submissions, sources and references are made available via github[8].

The collect submissions, we used the submission tool, OCELoT,[9] replacing the matrix that has been used up until 2019. Using OCELoT gives us more control over the submission and scoring process, for example we are able to limit the number of test submissions by each team, and we also display the submissions anonymously to avoid publishing any automatic scores.

For presentation of the results, systems are treated as either *constrained* or *unconstrained*. When the system submitters report that they were only trained on the provided data, we class them as constrained. The online systems are treated as unconstrained during the automatic and human evaluations, since we do not know how they were built.

In Appendix C, we provide brief details of the submitted systems, for those where the authors provided such details.

3 Human Evaluation

A human evaluation campaign is run each year to assess translation quality and to determine the official ranking of systems taking part in the news translation task. This section describes how data for the human evaluation is prepared, the process of collecting human assessments, and computation of the official results of the shared task.

3.1 Direct Assessment

We have employed Direct Assessment (DA, Graham et al., 2013, 2014, 2016) as the primary mechanism for evaluating systems since running a comparison of DA and relative ranking in 2016 (Bojar et al., 2016). DA has several important features including accurate quality control of crowdsourcing. With DA human evaluation, human assessors are asked to rate a given translation by how adequately it expresses the meaning of the corresponding reference translation or source language input on an analogue scale, which corresponds to an underlying absolute 0–100 rating scale.[10]

3.1.1 Source and Reference-based Evaluations

The original definition of DA provides human assessors with a reference translation. The benefit of this reference-based evaluation is that only speakers of the target language are needed, but the quality of the reference translation becomes critical and even if flawless, evaluating against a single reference translation could bias evaluators towards that reference.

In 2018, we trialled source-based (or "bilingual") evaluation for the first time, for English to Czech translation. In this configuration, the human assessor is shown the source input and system output only (with no reference translation shown). The assessor thus has to understand both the source and target languages very well but the quality of the reference is no longer vital. In fact, the human-generated reference can be included in the evaluation as an additional system to provide an estimate of human performance.

[6] https://archive.org/web/

[7] http://data.statmt.org/news-crawl

[8] https://github.com/wmt-conference/wmt21-news-systems

[9] https://github.com/AppraiseDev/OCELoT

[10] No sentence or document length restriction is applied during manual evaluation. Direct Assessment is also employed for evaluation of video captioning systems at TRECvid (Graham et al., 2018; Awad et al., 2019, 2021) and multilingual surface realisation (Mille et al., 2018, 2019).

Team	Language Pairs	System Description
AFRL	ru-en	(Erdmann et al., 2021)
ALLEGRO.EU	en-is,is-en	(Koszowski et al., 2021)
AMU	ha-en,en-ha	(Nowakowski and Dwojak, 2021)
BJTU-NMT	en-zh	(no associated paper)
BORDERLINE	en-zh,de-en,zh-en	(Wang et al., 2021)
BUPT-RUSH	en-zh,en-ja,en-de	(no associated paper)
CAPITALMARVEL	en-zh,en-ja,ja-en	(no associated paper)
CUNI-DOCTRANSFORMER	en-cs,cs-en	(Gebauer et al., 2021)
CUNI-MARIAN-BASELINES	en-cs	(Gebauer et al., 2021)
CUNI-TRANSFORMER2018	en-cs,cs-en	(Gebauer et al., 2021)
DIDI-NLP	zh-en	(no associated paper)
EPHEMERALER	en-zh,en-ja	(no associated paper)
ETRANSLATION	fr-de,en-cs,en-de	(Oravecz et al., 2021)
FACEBOOK-AI	ha-en,en-zh,en-ha,en-is,en-ja,de-en, zh-en,en-ru,en-cs,cs-en,ru-en,en-de, ja-en,is-en	(Tran et al., 2021)
FJDMATH	xh-zu	(Martinez, 2021)
GTCOM	ha-en,bn-hi,en-ha,zu-xh,hi-bn,xh-zu	(Bei and Zong, 2021)
HAPPYNEWYEAR	en-zh,zh-en	(no associated paper)
HAPPYPOET	en-zh,de-en,en-de	(no associated paper)
HW-TSC	ha-en,en-zh,bn-hi,en-ha,en-is,en-ja, zu-xh,de-en,zh-en,hi-bn,xh-zu,en-de, ja-en,is-en	(Wei et al., 2021)
ICL	en-zh,de-en,zh-en,en-de	(no associated paper)
IIE-MT	zh-en,ja-en	(no associated paper)
ILLINI	en-ja,ja-en	(Le et al., 2021)
KWAINLP	zh-en,ja-en	(no associated paper)
LAN-BRIDGE-MT	en-zh,en-is	(no associated paper)
LISN	fr-de,de-fr	(Xu et al., 2021)
MACHINE-TRANSLATION	en-zh,zh-en	(no associated paper)
MANIFOLD	ha-en,en-ha,en-is,de-en,en-ru,de-fr, ru-en,en-de,is-en	(no associated paper)
MIDEIND	en-is,is-en	(Jónsson et al., 2021)
MISS	en-zh,en-ja,zh-en,ja-en	(Li et al., 2021b)
MOVELIKEAJAGUAR	en-zh,en-ja,ja-en	(no associated paper)
MS-EGDC	ha-en,bn-hi,en-ha,zu-xh,hi-bn,xh-zu	(Hendy et al., 2021)
NIUTRANS	ha-en,en-zh,en-ha,en-is,en-ja,zh-en, en-ru,ru-en,ja-en,is-en	(Zhou et al., 2021)
NJUSC-TSC	en-zh,zh-en	(no associated paper)
NUCLEAR-TRANS	en-zh,en-de	(no associated paper)
NVIDIA-NEMO	de-en,en-ru,ru-en,en-de	(Subramanian et al., 2021)
P3AI	ha-en,en-zh,en-ha,fr-de,de-en,zh-en, de-fr,en-de	(Zhao et al., 2021)
SMU	en-zh,de-en,zh-en	(no associated paper)
TALP-UPC	fr-de,de-fr	(Escolano et al., 2021)
TRANSSION	ha-en,bn-hi,en-ha,zu-xh,hi-bn,xh-zu	(no associated paper)
TWB	ha-en,en-ha	(no associated paper)
UEDIN	ha-en,bn-hi,en-ha,de-en,hi-bn,en-de	(Chen et al., 2021; Pal et al., 2021)
UF	en-zh,de-en,zh-en,en-de	(no associated paper)
VOLCTRANS-AT	de-en,en-de	(Qian et al., 2021)
VOLCTRANS-GLAT	de-en,en-de	(Qian et al., 2021)
WATERMELON	de-en	(no associated paper)
WECHAT-AI	en-zh,en-ja,en-de,ja-en	(Zeng et al., 2021)
WINDFALL	en-zh	(no associated paper)
XMU	zh-en,ja-en	(no associated paper)
YYDS	en-zh,zh-en	(no associated paper)
ZENGHUIMT	en-zh,zh-en	(Zeng, 2021)
ZMT	ha-en,en-ha	(no associated paper)

Table 2: Participants in the shared translation task. The translations from the online systems were not submitted by their respective companies but were obtained by us, and are therefore anonymized in a fashion consistent with previous years of the workshop.

For both reference and source-based evaluation, we require human assessors to only evaluate translation *into* their native language. Following WMT19 and WMT20, we thus again use the source-based evaluation only for out-of-English language pairs. This is especially relevant since we have a large group of volunteer human assessors with native language fluency in non-English languages and high fluency in English, while we generally lack the reverse, i.e. native English speakers with high fluency in non-English languages.

We use different implementation and human annotators for into-English and out-of-English. We describe the approaches separately. Reference-based (monolingual) into-English human evaluation is described in Section 3.2, while source-based (bilingual) out-of-English and non-English human evaluation is described in Section 3.3. A third, simplified annotation was used for Bengali↔Hindi and Xhosa↔Zulu, Section 3.4.

3.1.2 Translationese

Prior to WMT19, all the test sets included a mix of sentence pairs that were originally in the source language, and then translated to the target language, and sentence pairs that were originally in the target language but translated to the source language. The inclusion of the latter "reverse-created" sentence pairs has been shown to introduce biases into the evaluations, particularly in terms of BLEU scores (Graham et al., 2020). Therefore we have avoided it for all language pairs, apart from Bengali↔Hindi and Xhosa↔Zulu, where the texts are all translated from English.

3.1.3 Document Context

As mentioned already in our discussion in WMT18 and as also established within the community (Läubli et al., 2018b; Toral et al., 2018a), evaluating sentences out of their document context can skew the results. The effect is particularly pronounced when comparing human and machine translation, where it is observed that evaluators tend to rate the human translation higher (relative to the machine translation) when the translations are viewed in context. Human translators always have access to the document context when translating to create the references.

In WMT19, we experimented with a DA style that considers document context in a simple way.

Language Pair	Sys.	Assess.	Assess/Sys
Czech→English	9	10,651	1,183.4
German→English	20	25,718	1,285.9
Hausa→English	14	17,321	1,237.2
Icelandic→English	10	11,124	1,112.4
Japanese→English	16	17,055	1,065.9
Russian→English	11	11,499	1,045.4
Chinese→English	24	44,268	1,844.5
Total to-English	104	137,636	1,323.4

Table 3: Amount of data collected in the WMT21 manual evaluation campaign for evaluation into-English; after removal of quality control items.

Dubbed "SR+DC" (segment rating with document context), this method presents one segment at a time but the segments are no longer shuffled (as in "SR−DC", segment rating without document context). Instead, they are provided in the order in which they appear in the document. The implementation still has the limitation that the assessors cannot go back to the previous segment.

An improved alternative to "SR+DC" is to offer the full document and allow the assessors to review their segment-level ratings. We call this setup "SR+FD" (segment ranking in a full document) and illustrate the user interface in Appraise in Figure 5.[11]

This year, for all language pairs for which document context was available, we include it when evaluating translations. Note that the ratings are nevertheless collected on the segment level, motivated by the power analysis described in Graham et al. (2019) and Graham et al. (2020). The particular details on how document context is made available to assessors depends on the translation direction, as described in more detail in Sections 3.2 to 3.4.

3.2 Human Evaluation of Translation into-English

In terms of the News translation task manual evaluation for into-English language pairs, a total of 589 turker accounts were involved.[12] 488,396 translation assessment scores were submitted in total by the crowd, of which 170,194 were provided by workers who passed quality control.[13]

System rankings are produced from a large set of human assessments of translations, each of which indicates the absolute quality of the out-

[11]Compare with Figures 3 and 4 in Bojar et al. (2019).

[12]Numbers do not include the 1,078 workers on Mechanical Turk who did not pass quality control.

[13]Numbers include quality control segments.

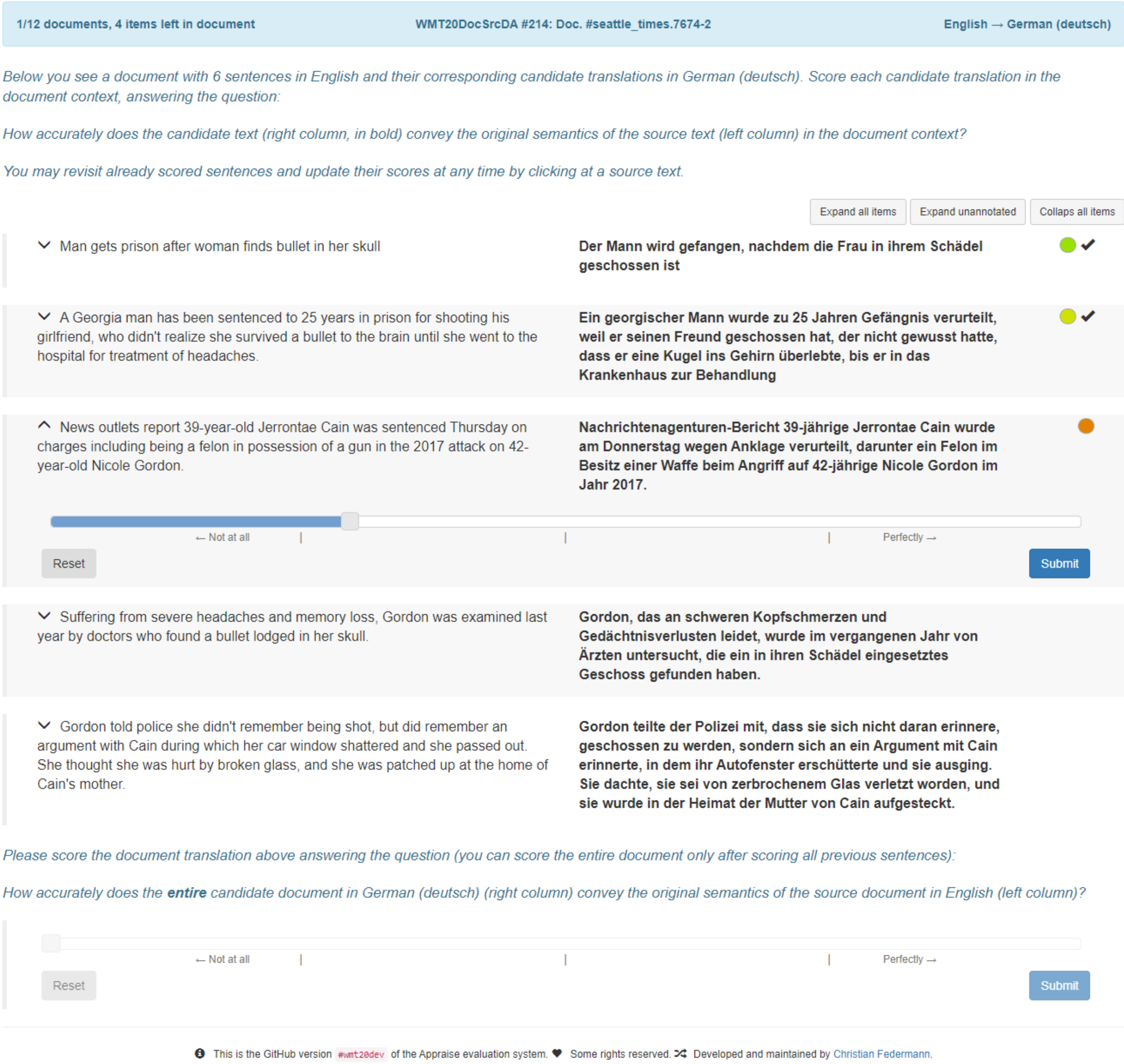


Figure 5: Screen shot of the document-level DA (SR+FD, segment rating within the full document) configuration in the Appraise interface for an example assessment from the human evaluation campaign. The annotator is presented with the entire translated document randomly selected from competing systems (anonymized) and is asked to rate the translation of individual segments and then entire document on sliding scales.

put of a system. Table 3 shows total numbers of human assessments collected in WMT21 for into-English language pairs contributing to final scores for systems.[14]

3.2.1 Crowd Quality Control

Collection of segment-level ratings with document context (SR+DC, Segment Rating + Document Context) involved constructing HITs so that each sentence belonging to a given document (produced by a single MT system) was displayed to and rated in turn by the human annotator.

We then injected the three kinds of quality control translation pairs described in Table 4: we repeat pairs expecting a similar judgment (Repeat Pairs), damage MT outputs expecting significantly worse scores (Bad Reference Pairs) and use references instead of MT outputs expecting high scores (Good Reference Pairs). For each of these three types, we include the MT output, along with its corresponding control item.

HITs were then constructed as follows, with as close as possible to 100 segments in a single HIT:

1. All documents produced by all systems are pooled;[15]

[14] Number of systems for WMT21 includes four "human" systems comprising human-generated reference translations used to provide human performance estimates.

[15] If a "human" system is included to provide a human per-

Repeat Pairs:	Original System output (10)	An exact repeat of it (10);
Bad Reference Pairs:	Original System output (10)	A degraded version of it (10);
Good Reference Pairs:	Original System output (10)	Its corresponding reference translation (10).

Table 4: Standard DA HIT structure quality control translation pairs hidden within 100-translation HITs, numbers of items are provided in parentheses.

2. Documents are then sampled at random (without replacement) and assigned to the current HIT until the current HIT contains close to (but less than) 70 segments

3. Once documents amounting to close to 70 segments have been assigned to the current HIT, we select a subset of these documents to be paired with quality control documents; this subset is selected by repeatedly checking if the addition of the number of the segments belonging to a given document (as quality control items) will keep the total number of segments in the HIT below 100; if this is the case, it is included; otherwise it is skipped until the addition of all documents has been checked. In doing this, the HIT is structured to bring the total number of segments as close as possible to 100 segments.

4. Once we have selected a core set of original system output documents and a subset of them to be paired with quality control versions for each HIT, quality control documents are automatically constructed by altering the sentences of a given document into a mixture of three kinds of quality control items used in the original DA segment-level quality control: bad reference translations, reference translations and exact repeats (see below for details of bad reference generation and Table 5 for numbers of words replaced in document segments);

5. Finally, the documents belonging to a HIT are shuffled.

Construction of Bad References As in previous years, bad reference pairs were created automatically by replacing a phrase within a given translation with a phrase of the same length, randomly selected from n-grams extracted from the full test set of reference translations belonging to that language pair. This means that the replacement phrase will itself comprise a mostly fluent sequence of words (making it difficult to tell that the sentence is low quality without reading the entire sentence) while at the same time making its presence highly likely to sufficiently change the meaning of the MT output so that it causes a noticeable degradation. The length of the phrase to be replaced is determined by the number of words in the original translation, as listed in Table 5.

Translation Length (N)	# Words Replaced in Translation
1	1
2–5	2
6–8	3
9–15	4
16–20	5
>20	$\lfloor$ N/4 $\rfloor$

Table 5: Number of words replaced when constructing quality control items.

Quality Filtering When an analogue scale (or 0–100 point scale, in practice) is employed, agreement cannot be measured using the conventional Kappa coefficient, ordinarily applied to human assessment when judgments are discrete categories or preferences. Instead, to measure consistency we filter crowd-sourced human assessors by how consistently they rate translations of known distinct quality using the bad reference pairs described previously. Quality filtering via bad reference pairs is especially important for the crowd-sourced portion of the manual evaluation. Due to the anonymous nature of crowd-sourcing, when collecting assessments of translations, it is likely to encounter workers who attempt to game the service, as well as submission of inconsistent evaluations and even robotic ones. We therefore employ DA's quality control mechanism to filter out low quality data, facilitated by the use of DA's analogue rating scale.

Assessments belonging to a given crowd-source worker who has not demonstrated that he/she can reliably score bad reference translations significantly lower than corresponding genuine system

formance estimate, it is also considered a system during quality control set-up.

	All	(A) Sig. Diff. Bad Ref.	(A) & No Sig. Diff. Exact Rep.
Czech→English	290	73 (25%)	68 (93%)
German→English	605	162 (27%)	150 (93%)
Hausa→English	423	109 (26%)	101 (93%)
Icelandic→English	273	75 (27%)	67 (89%)
Japanese→English	315	103 (33%)	91 (88%)
Russian→English	187	84 (45%)	77 (92%)
Chinese→English	617	195 (32%)	178 (91%)
Total	**1,694**	**589 (35%)**	**544 (92%)**

Table 6: Number of crowd-sourced workers taking part in the reference-based SR+DC campaign; (A) those whose scores for bad reference items were significantly lower than corresponding MT outputs; those of (A) whose scores also showed no significant difference for exact repeats of the same translation; note: many workers evaluated more than one language pair.

output translations are filtered out. A paired significance test is applied to test if degraded translations are consistently scored lower than their original counterparts and the p-value produced by this test is used as an estimate of human assessor reliability. Assessments of workers whose p-value does not fall below the conventional 0.05 threshold are omitted from the evaluation of systems, since they do not reliably score degraded translations lower than corresponding MT output translations.

Table 6 shows the number of workers participating in the into-English translation evaluation who met our filtering requirement in WMT21 by showing a significantly lower score for bad reference items compared to corresponding MT outputs, and the proportion of those who simultaneously showed no significant difference in scores they gave to pairs of identical translations. We removed data from the non-reliable workers in all language pairs.

3.2.2 Producing the Human Ranking

This year all rankings (for to-English translation) were arrived at via segment ratings presented one at a time in their original document order (SR+DC).

In order to iron out differences in scoring strategies of distinct human assessors, human assessment scores for translations were first standardized according to each individual human assessor's overall mean and standard deviation score.

Average standardized scores for individual segments belonging to a given system were then computed, before the final overall DA score for a given system is computed as the average of its segment scores (Ave z in Table 7). Results are also reported for average scores for systems, computed in the same way but without any score standardization applied (Ave % in Table 7).

Human performance estimates arrived at by evaluation of human-produced reference translations are denoted by "HUMAN" in all tables.

Clusters are identified by grouping systems together according to which systems significantly outperform all others in lower ranking clusters, according to Wilcoxon rank-sum test. Rank ranges are based on the same head-to-head statistical significance tests. For instance, if a system is statistically significantly worse than 2 other systems, and not statistically different from 4 other systems, its rank is reported as 3–6 (the top of the rank range is 2+1, the bottom 2+4).

All data collected during the human evaluation is available at `http://www.statmt.org/wmt21/results.html`. Appendix A shows the underlying head-to-head significance test official results for all pairs of systems and also reports BLEU, chrF, and COMET scores.

3.3 Bilingual Human Evaluation

Human evaluation for nine out-of-English and non-English translation directions used a source-based (sometimes called "bilingual") direct assessment of individual segments in the full document context (SR+FD), as established in WMT20 (Barrault et al., 2020).

In an attempt to break more ties among the participating systems, we also ran a second stage of annotation using segment-level contrastive source-based DA ignoring document context (labelled "contr:SR−DC") for top-10 systems (plus human references) for 3 out-of-English language pairs. Details on the second stage are in Section 3.3.5.

In the source-based DA campaign, we collected 303,627 assessments in total after excluding quality control items and users who did not pass the quality control. The contrastive source-based DA campaign provided 64,031 translation assessments. The total numbers of collected assessments per language pair are presented in Table 8. For data collection, we used the open-source Appraise Evaluation Framework (Federmann, 2012) for both assessment types.

Czech→English

Rank	Ave.	Ave. z	System
1–2	77.8	0.111	Facebook-AI
1–2	78.4	0.081	Online-A
3–6	72.0	0.008	CUNI-DocTransf
3–6	74.0	−0.005	Online-B
3–8	71.5	−0.008	CUNI-Trf2018
3–8	74.5	−0.032	Online-W
5–9	67.2	−0.039	Online-G
7–9	74.4	−0.084	Online-Y
5–9	75.6	−0.085	HUMAN-B

German→English

Rank	Ave.	Ave. z	System
1–5	71.9	0.126	Borderline
1–6	73.5	0.124	Online-A
1–4	78.6	0.122	Online-W
4	79.5	0.113	UF
3–8	73.2	0.106	VolcTrans-AT
4–9	77.5	0.100	Facebook-AI
5–12	75.8	0.068	ICL
4–12	73.4	0.048	Online-G
8–17	69.7	0.016	Online-B
7–17	71.3	0.016	Online-Y
7–17	71.6	0.010	VolcTrans-GLAT
5–16	69.6	0.007	P3AI
9–19	70.6	−0.008	SMU
9–17	73.1	−0.008	UEdin
9–17	69.1	−0.010	NVIDIA-NeMo
10–19	69.9	−0.035	Manifold
15–20	67.0	−0.043	Watermelon
7–17	71.8	−0.061	happypoet
16–20	66.8	−0.081	HUMAN-C
18–20	66.0	−0.120	HW-TSC

Hausa→English

Rank	Ave.	Ave. z	System
1	74.4	0.248	Facebook-AI
2–4	68.8	0.118	Online-B
3–7	66.6	0.062	TRANSSION
2–6	66.5	0.059	ZMT
3–6	69.0	0.059	GTCOM
3–9	65.3	0.029	HW-TSC
5–19	65.2	0.002	MS-EgDC
6–10	60.1	−0.031	P3AI
6–10	62.4	−0.032	NiuTrans
8–11	63.5	−0.090	Online-Y
10–12	59.6	−0.112	Manifold
11–13	60.4	−0.173	AMU
12–13	58.2	−0.205	UEdin
14	56.9	−0.267	TWB

Icelandic→English

Rank	Ave.	Ave. z	System
1	74.5	0.293	Facebook-AI
2	74.8	0.112	Manifold
3–7	75.1	0.045	NiuTrans
3–8	71.3	0.028	Online-B
3–7	76.6	0.013	HW-TSC
3–7	69.7	0.009	Mideind
3–9	75.4	0.003	Online-A
6–9	70.1	−0.037	Allegro.eu
7–9	71.7	−0.080	Online-Y
10	65.2	−0.256	Online-G

Japanese→English

Rank	Ave.	Ave. z	System
1	73.8	0.141	HW-TSC
2–5	65.1	0.082	IIE-MT
2–6	68.6	0.046	NiuTrans
2–9	67.8	0.033	KwaiNLP
2–6	66.2	0.032	Facebook-AI
5–11	63.5	0.025	XMU
3–10	66.8	0.011	capitalmarvel
5–11	60.9	0.001	Online-B
6–11	61.5	−0.031	MiSS
5–11	66.7	−0.039	Online-W
7–12	59.3	−0.062	WeChat-AI
11–14	59.0	−0.080	Online-A
12–16	55.0	−0.140	Online-G
12–16	64.8	−0.157	movelikeajaguar
13–16	62.2	−0.189	Online-Y
13–16	55.4	−0.193	Illini

Russian→English

Rank	Ave.	Ave. z	System
1–5	77.5	0.137	NVIDIA-NeMo
1–4	73.9	0.130	Online-W
3–7	73.1	0.108	Online-B
1–7	73.3	0.089	HUMAN-B
2–7	71.7	0.060	Manifold
1–7	70.4	0.056	Facebook-AI
3–8	68.5	0.044	NiuTrans
7–10	65.1	0.016	Online-G
8–11	65.5	−0.014	AFRL
8–11	63.9	−0.022	Online-A
9–12	69.1	−0.123	Online-Y

Chinese→English

Rank	Ave.	Ave. z	System
1–5	75.0	0.042	NiuTrans
1–6	77.0	0.039	KwaiNLP
1–6	75.6	0.031	DIDI-NLP
1–9	74.1	0.019	HUMAN-B
1–9	71.7	0.016	HappyNewYear
2–19	74.0	−0.001	P3AI
4–18	70.5	−0.023	Borderline
4–19	72.6	−0.026	ICL
6–17	70.1	−0.029	MiSS
3–24	73.1	−0.031	IIE-MT
9–22	72.8	−0.032	Machine-Translation
7–21	70.6	−0.034	SMU
7–21	70.7	−0.036	yyds
6–20	70.1	−0.037	Facebook-AI
7–21	73.6	−0.042	Online-B
7–21	73.5	−0.050	ZengHuiMT
7–21	73.0	−0.062	HW-TSC
7–22	67.6	−0.068	XMU
12–24	76.0	−0.072	NJUSC-TSC
11–24	72.1	−0.082	Online-G
8–22	72.9	−0.087	Online-W
17–24	70.1	−0.103	UF
20–24	66.7	−0.106	Online-A
20–24	69.0	−0.174	Online-Y

Table 7: Official results of WMT21 News Translation Task for translation into-English (SR+DC). Systems ordered by DA score z-score; systems within a cluster are considered tied; lines indicate clusters according to Wilcoxon rank-sum test $p < 0.05$; rank ranges are based on the same test (for details, see Section 3.2.2); grayed entry indicates resources that fall outside the constraints provided.

Language Pair	Sys.	Assess.	Assess/Sys
English-Czech	12	50,491	4,207.6
English-German	22	24,689	1,122.2
English-Hausa	15	18,656	1,243.7
English-Icelandic	12	16,940	1,411.7
English-Japanese	16	43,991	2,749.4
English-Russian	11	31,632	2,875.6
English-Chinese	31	84,322	2,720.1
German-French	10	21,018	2,101.8
French-German	10	11,888	1,188.8
Total standard DA	139	303,627	2,184.4
English-Czech	12	19,279	1,606.6
English-German	12	23,212	1,934.3
English-Chinese	12	21,540	1,795.0
Total contrastive DA	36	64,031	1,778.6

Table 8: Amount of data collected in the WMT21 manual document- and segment-level evaluation campaigns for bilingual source-based evaluation out-of-English and non-English language pairs. The system counts include the human references (either 1 or 2 references, depending on language pair).

3.3.1 Sources of Human Annotators

We used three groups of annotators: participants in the News Shared Task, crowd-workers from the Toloka platform, and paid professional annotators sponsored by Microsoft.

We asked participants of the news task to contribute around 9 hours of annotation time (which we estimated at 12 HITs) per each primary system submitted, with each HIT including roughly 100 segment translations. Furthermore, we collected information about the classification of their annotators type. Unfortunately, only 65% of the requested annotations were finished by participating teams.

The second annotator group was provided by Toloka AI.[16] Toloka AI is a global data labeling company that helps its customers generate machine learning data at scale by harnessing the wisdom of the crowd from around the world. It relies on a geographically diverse crowd of several million registered users (Pavlichenko et al., 2021).[17] Toloka tests proficiency of their annotator crowd and excludes from future annotations anyone who does not pass quality control in the Appraise tool.

The last part of annotations is sponsored by Microsoft, who contributed with their crowd of qualified paid bilingual speakers experienced in the annotation process. Moreover, Microsoft tracks the performance of the annotators, and those who fail quality control are permanently removed from the pool of annotators. This increases the overall quality of the human assessment.

For bilingual human evaluation, Microsoft contributed with 42%, WMT News participants contributed with 37%, and Toloka platform with 21% of all valid annotations (after removal of annotators that do not pass quality control). The distribution of individual groups of annotators per each language is presented in Table 9.

3.3.2 Document-Level Assessment

This year's human evaluation for out-of-English and non-English language pairs features a document-level direct assessment configuration as presented last year (Barrault et al., 2020). We again use the segment level rating but provide the full document at once (SR+FD, segment rating within a full document), for a more reliable evaluation (Castilho et al., 2020; Laubli et al., 2020).

Figure 5 above shows a screenshot of the fully document-level interface. In the default scenario, an annotator scores individual segments one by one and, after scoring all of them, on the same screen, the annotator then judges the translation of the entire document displayed. Annotators can, however, revisit and update scores of previously assessed segments at any point of the annotation of the given document. It has been shown that presenting the entire document context on a screen may lead to higher quality segment- and document-level assessments (Grundkiewicz et al., 2021) improving the correlation between segment and document scores and increasing inter-annotator agreement for document scores. A similar setup has been used by Popel et al. (2020) even for more than two systems compared at once.

3.3.3 Quality Control

For the document-level evaluation of out-of-English translations, HITs were generated using the same method as described for the SR+DC evaluation of into-English translations in Section 3.2.1 with a minor modification: Since the annotations are made by researchers and professional translators who ensure a better quality of assessments than the crowd-sourced workers, only bad references are used as quality control items.

[16] https://toloka.ai/
[17] https://hackernoon.com/evolution-of-the-data-production-paradigm-in-ai

	Microsoft annotators	Toloka paid crowd	Participants linguists	annotators	researchers	students
English - Chinese	33%	11%	2%	20%	17%	17%
English - Czech	27%	18%	-	54%	-	-
English - German	56%	29%	13%	-	2%	-
English - Hausa	63%	35%	3%	-	-	-
English - Icelandic	82%	5%	13%	-	-	-
English - Japanese	43%	20%	1%	26%	4%	8%
English - Russian	29%	39%	9%	-	23%	-
French - German	76%	14%	11%	-	-	-
German - French	43%	45%	11%	-	-	-
Total	42%	21%	37%			

Table 9: Distribution of annotation crowds for each language pair in bilingual human evaluation. Annotator types are self-classified by participants.

3.3.4 Including Human Translations

Source-based DA allows us to include human references in the evaluation as another system to provide an estimate of human performance. Human references were added to the pool of system outputs prior to sampling documents for tasks generation. Each reference is assessed individually if multiple references are available, which is the case for English→German, English→Czech, English→Russian, and English→Chinese.

3.3.5 Contrastive Direct Assessment

This year we extended the bilingual source-based human evaluation with contrastive evaluation using segment-level pairwise direct assessments (Novikova et al., 2018; Sakaguchi and Van Durme, 2018). It has been pointed out (Freitag et al., 2021a) that standard direct assessment may not be able to properly differentiate high-quality MT system outputs. The contrastive approach to DA can strengthen the discriminative power as annotators judge translations in relation to each other. When standard DA can likely provide better *absolute* quality assessment, the contrastive evaluation can provide better *relative* quality assessments between system pairs. This may help create a more reliable ranking of systems if used on top of the standard approach described in Section 3.3.

The contrastive evaluation is similar to the relative ranking used from WMT08 (Callison-Burch et al., 2008) to WMT16 (Bojar et al., 2016), where annotators were presented with up to five system outputs and corresponding source and reference sentence and asked to rank these systems between each other. The main differences in this year's contrastive evaluation to the relative rankings are that 1) the evaluation is source-based, i.e. without the reference, 2) the continuous scale is used instead of ranks, and 3) only two system outputs are judged at the same time instead of five.

To reduce the cognitive load on annotators, we decided to trial this contrastive approach evaluating individual sentences independent of their context. This is a very important difference compared to the the first stage (Section 3.3).

We ran the contrastive evaluation for English→Chinese, English→Czech and English→German, and we selected top-10 best performing systems based on DA z-score from the ranking created using standard direct assessment for those languages (Table 10), and two human references.

This contrastive evaluation was sponsored by Microsoft and performed by the bilingual paid annotator group as described in Section 3.3.1. Assessments were collected using the open-source Appraise Evaluation Framework (Federmann, 2012). A screenshot of the user interface used in this stage is shown in Figure 6. Each annotator is presented with two randomly selected translated segments from competing systems (anonymized) and asked to rate both of them on a continuous scale of 0-100. Upon request by the annotator, the differences between the two translations were highlighted at the word level to help avoid missing differences. This highlighting may however reduced the effectiveness of control items.

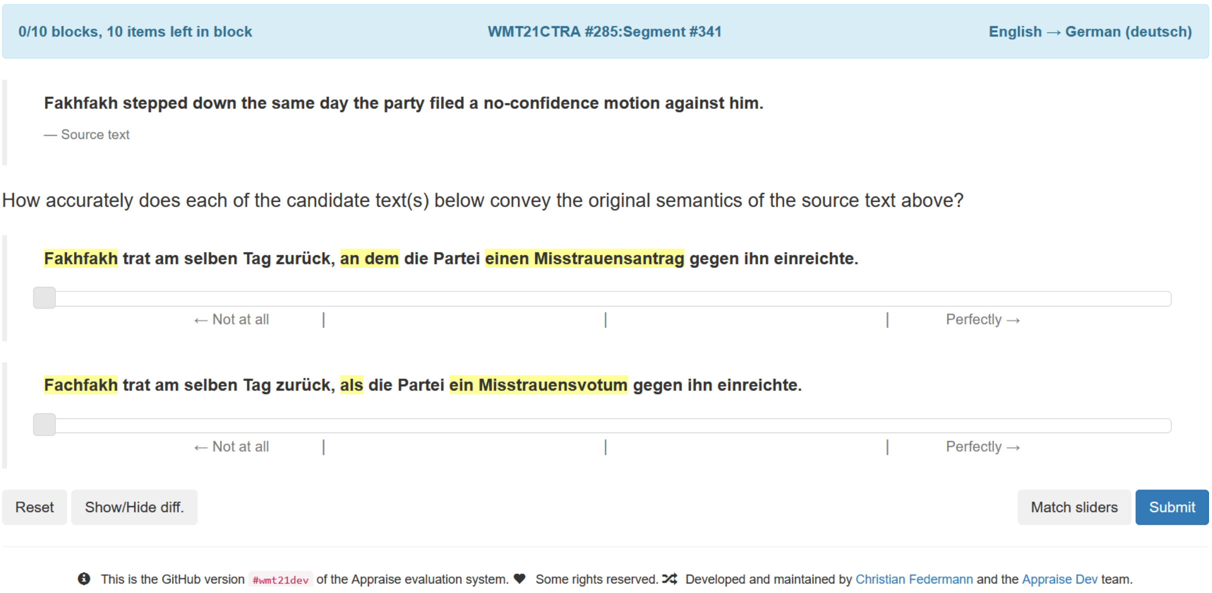


Figure 6: Screen shot of the contrastive DA configuration in the Appraise interface for an example assessment from the 2nd stage of human evaluation campaign. The annotator is presented with two translated segments randomly selected from competing system outputs (anonymized) and is asked to rate both of them on sliding scales.

3.3.6 Human Rankings

Table 10 shows official news task results for translation out-of-English, where lines indicate clusters according to Wilcoxon rank-sum test $p < 0.05$.

Source-based DA scores were collected based on the document-level annotation interface, so context was available during annotation. All systems are evaluated in isolation, based on the annotators' perception of translation quality given the source text and document context. Across all language pairs, human reference translations end up in the top-scoring cluster, indicative of a (relatively) high quality of these references. For language pairs with large numbers of submissions, we observe little to no clustering. Notably English→German has only two clusters, one of which contains all but one of the submitted systems, and English→Chinese ends up with a huge mono cluster containing all submissions. While there are differences in average scores and z scores these are not statistically significant enough for effective clustering. As a substitute, rank ranges give an indication of the respective system's translation quality.

Table 11 shows contrastive news task results for translation out-of-English, where lines indicate clusters according to Wilcoxon rank-sum test $p < 0.05$.

Contrastive, source-based DA scores (contr:SR−DC) were collected using a segment-level annotation interface, so context was *not* been available to annotators. Results for the source-based DA annotation phase (SR+FD) in Table 11 were computed on the subset of data for the ten systems and two references for which we have run the contrastive, source-based DA annotation phase.

We generally observe better clustering for the contr:SR−DC. This is especially noteworthy as the number of annotations collected per system is much higher for the first, SR+FD, DA phase (for two of the three language pairs on which contr:SR−DC was run). It seems that pairwise comparison of system outputs is beneficial for determining whether differences between systems are statistically significant.

In contrast to the first annotation phase, we find that human reference translations are scored worse, and significantly worse than the top cluster. We explain this by the fact that our contrastive setup was run on segment-level while the source-based DA annotators had access to the full document context. A simple explanation that should nevertheless be empirically validated is that the wording of the sentence created for and within the context of the document does not sound flawless and natural when evaluated in isolation (Läubli et al., 2018a; Toral et al., 2018b). Some machine translation systems do consider the surrounding sentences but their capacity of 'contextualizing' the candidate sentences is probably limited.

Observing the striking difference in system

English→Czech

Rank	Ave.	Ave. z	System
1	90.2	0.397	HUMAN-A
2–4	87.9	0.284	HUMAN-B
2–4	87.6	0.263	Facebook-AI
2–4	86.1	0.214	Online-W
5–7	83.0	0.122	eTranslation
5–6	82.1	0.047	CUNI-Transformer2018
6–8	79.2	-0.120	CUNI-DocTransformer
7–9	79.3	-0.154	CUNI-Marian-Baselines
8–10	77.8	-0.183	Online-B
9–10	74.6	-0.308	Online-A
11	76.2	-0.373	Online-Y
12	65.6	-0.674	Online-G

English→Icelandic

Rank	Ave.	Ave. z	System
1	88.1	0.872	HUMAN-A
2	84.5	0.594	Facebook-AI
3–4	68.2	0.277	NiuTrans
3–4	72.7	0.240	Manifold
5–9	75.2	0.200	Online-A
5–7	65.6	0.130	Lan-Bridge-MT
5–9	62.6	0.063	Mideind
6–9	73.9	0.026	Online-B
6–9	75.6	-0.034	HW-TSC
10	62.0	-0.236	Online-Y
11	48.7	-0.470	Allegro.eu
12	33.9	-1.082	Online-G

English→Chinese

Rank	Ave.	Ave. z	System
1–3	82.5	0.325	HUMAN-B
2–14	74.9	0.284	HappyNewYear
1–7	81.2	0.250	Facebook-AI
1–8	80.0	0.216	HUMAN-A
4–19	75.3	0.164	Borderline
2–19	81.0	0.161	bjtu_nmt
3–14	75.5	0.151	Lan-Bridge-MT
4–21	79.3	0.124	BUPT_rush
2–18	79.2	0.098	NiuTrans
4–18	75.7	0.091	Machine_Translation
2–15	80.9	0.078	SMU
6–22	81.4	0.064	capitalmarvel
4–19	79.5	0.056	WeChat-AI
6–22	78.1	0.026	Online-W
7–22	75.2	0.004	ICL
9–23	75.9	-0.008	HW-TSC
5–23	78.2	-0.025	ZengHuiMT
11–22	81.2	-0.026	yyds
10–26	79.7	-0.050	P3AI
17–27	77.1	-0.061	windfall
6–24	78.9	-0.075	Online-B
13–26	76.8	-0.080	NJUSC_TSC
9–24	77.7	-0.100	MiSS
19–27	77.0	-0.101	UF
22–28	72.7	-0.123	Online-A
22–28	79.3	-0.160	happypoet
20–28	76.9	-0.185	nuclear_trans
25–29	76.4	-0.247	ephemeraler
28–31	67.5	-0.257	Online-G
29–31	67.1	-0.463	Online-Y
29–31	68.3	-0.613	movelikeajaguar

English→German

Rank	Ave.	Ave. z	System
1–17	83.3	0.266	Online-B
1–5	84.7	0.243	Online-W
1–14	86.6	0.217	WeChat-AI
1–6	87.6	0.145	Facebook-AI
1–10	89.4	0.116	UF
2–17	85.2	0.089	HW-TSC
3–17	86.8	0.072	UEdin
3–18	86.5	0.041	P3AI
3–18	86.4	0.030	HUMAN-A
5–19	83.3	0.013	happypoet
4–19	86.1	0.010	eTranslation
4–19	84.4	0.001	Online-A
3–18	84.5	0.001	HUMAN-C
5–19	78.8	-0.053	VolcTrans-AT
5–19	86.7	-0.055	NVIDIA-NeMo
8–21	83.1	-0.058	Manifold
4–20	84.3	-0.062	Online-G
12–20	84.5	-0.072	Online-Y
18–21	73.9	-0.130	ICL
4–20	85.0	-0.140	VolcTrans-GLAT
16–21	78.3	-0.179	nuclear_trans
22	80.0	-0.415	BUPT_rush

English→Japanese

Rank	Ave.	Ave. z	System
1–2	86.4	0.430	Facebook-AI
1–2	85.3	0.314	HUMAN-A
3–5	84.2	0.266	Online-W
3–5	81.3	0.168	WeChat-AI
3–5	82.6	0.148	NiuTrans
6–8	77.8	0.017	HW-TSC
6–8	71.8	-0.042	MiSS
8–13	78.5	-0.051	Online-Y
6–10	77.8	-0.067	BUPT_rush
8–13	70.9	-0.129	Online-A
9–13	67.4	-0.184	Online-B
9–14	74.2	-0.284	ephemeraler
9–14	72.5	-0.339	capitalmarvel
12–14	70.1	-0.373	movelikeajaguar
15–16	63.5	-0.440	Illini
15–16	65.7	-0.541	Online-G

French→German

Rank	Ave.	Ave. z	System
1–5	87.7	0.088	Online-W
1–7	89.2	0.052	Online-A
1–4	89.5	0.035	HUMAN-A
2–8	85.7	0.002	LISN
1–8	86.9	-0.014	Online-B
4–10	85.0	-0.021	talp_upc
3–8	85.0	-0.064	eTranslation
7–10	84.1	-0.154	Online-G
3–10	86.6	-0.210	Online-Y
7–10	86.4	-0.229	P3AI

English→Russian

Rank	Ave.	Ave. z	System
1–3	86.0	0.317	HUMAN-B
1–3	83.3	0.277	Online-W
1–3	82.5	0.093	HUMAN-A
4–6	79.4	0.056	Online-B
4–7	75.3	0.032	Online-A
4–7	80.1	-0.001	Facebook-AI
7–10	74.5	-0.123	NiuTrans
7–10	72.3	-0.153	Manifold
7–10	75.4	-0.161	NVIDIA-NeMo
5–10	76.0	-0.180	Online-G
11	62.7	-0.541	Online-Y

English→Hausa

Rank	Ave.	Ave. z	System
1–2	84.1	0.362	HUMAN-A
1–4	82.7	0.264	Facebook-AI
2–5	80.8	0.263	NiuTrans
3–6	81.2	0.175	Online-B
4–6	80.1	0.128	TRANSSION
2–6	79.2	0.124	ZMT
7–10	78.0	0.018	P3AI
7–10	78.7	0.006	HW-TSC
8–12	75.2	-0.026	AMU
7–10	78.8	-0.036	GTCOM
9–12	75.0	-0.128	MS-EgDC
12–15	70.2	-0.227	UEdin
11-15	73.4	-0.243	Manifold
12–15	70.5	-0.340	TWB
11-15	67.7	-0.448	Online-Y

German→French

Rank	Ave.	Ave. z	System
1–3	87.9	0.160	Online-B
1–3	86.5	0.126	HUMAN-A
3–6	83.4	0.018	Manifold
1–6	84.8	0.006	Online-W
3–6	84.5	0.004	Online-A
6–10	83.0	-0.084	Online-G
3–10	83.5	-0.148	P3AI
6–10	81.3	-0.149	LISN
6–10	83.7	-0.177	Online-Y
6–10	81.0	-0.190	talp_upc

Table 10: Official results of WMT21 News Translation Task for translation out-of-English (SR+FD). Systems ordered by DA score z-score; systems within a cluster are considered tied; lines indicate clusters according to Wilcoxon rank-sum test $p < 0.05$; rank ranges are based on the same test (for details, see Section 3.2.2); grayed entry indicates resources that fall outside the constraints provided. DA scores are collected using a document-level annotation interface, so context is available to annotators.

ranking by SR+FD vs. contr:SR−DC, esp. the discrepancy in the ranking of human translations, we conclude that evaluating MT systems without document context is no longer reliable for mid- and high-quality MT systems. This is also supported by the surprising observation in Czech→English in Table 7 where humans seemed to be surpassed by *all* participating MT systems. (Considering statistical significance, the claim is arguably weaker: humans share the second cluster with the majority of the systems.) We acknowledge that it is possible that the Czech→English HUMAN-B references are of much worse quality than the English→Czech ones,[18] but we tend to put more trust in the reference quality than in the SR+DC method for two reasons: (1) The annotators did not see the whole document at once and cannot go back in their annotation, so their effective capability to consider context is limited. (2) It is possible that other effects of reference-based DA in the Czech→English start playing role when both the candidate and reference are human vs. when only the reference is human. One possibility would be a stronger confidence of assessors when scoring human translations, leading e.g. to more polarized scores. A detailed investigation into manual evaluation methods that word reliably for both human and machine translations is thus still needed.

3.4 Human Evaluation of Bengali↔Hindi and Xhosa↔Zulu Translation (Wikipedia Data)

Translation quality for Bengali↔Hindi and Xhosa↔Zulu was evaluated using Direct Assessment without considering document context (SR−DC) with a scoring scale of 1-100 by vetted human evaluators. The human evaluators were asked to provide a judgment that they felt most accurately reflected the perceived quality of each corresponding translation of the give source sentence. Definitions of translation quality within several scoring ranges were provided to assist evaluators in providing consistent annotations.

A participating system translation was displayed on the right next to its corresponding source sentence on the left. The sentence pairs were then randomized and passed to a human evaluator for a single direct assessment. The evaluation was performed on the sentence level and evaluators provided a direct assessment score for each sentence-translation pair. The user interface was simpler than the one shown in Figure 5: instead of a slider, the annotators had to enter the scores numerically.

Because evaluators were extremely difficult to recruit for these language pairs and the evaluation was thus low resource, no quality control items were injected and we focused on the vetting process of the evaluators prior to performing any assessment. The only sanity check was that evaluators enter an integer between 1 and 100 as the scores.

All segments from the FLORES Wikimedia test set were included for the evaluation. Each segment was annotated and assessed by one evaluator only once.

All four language directions were assessed by trusted evaluators who have been vetted by a localization vendor specializing in translation evaluation services, to have native fluency of the target language, fluent to native understanding of the source language, have lived in the target region for at least five years recently, and have had at least two to five years of professional translation experience. For Hindi→Bengali and Bengali→Hindi, two human evaluators were used with the translation data being split in half and randomly assigned to the respective evaluators. Two human evaluators assessed for Xhosa→Zulu data and one evaluator assessed for Zulu→Xhosa. The number of evaluators and judgments they made is provided in Table 12.

The final scores for Bengali↔Hindi and Xhosa↔Zulu are provided in Table 13.

3.5 GENIE DE-EN Evaluation

This year, human evaluations for German→English translation with the GENIE leaderboard were also carried out. GENIE is an ongoing effort that centralizes and facilitates human evaluations for natural language generation tasks (Khashabi et al., 2021). In

[18] The quality assurance for each of "A" and "B" references for English↔Czech was comparable; not that the same translators would be producing both directions. In fact, we expected the "B" translations to be *better*, because they were created by experienced students and teachers of translation studies, who are active translators themselves and who *specifically attempted to produce as good translations as possible*. As the to-Czech scores suggest, our annotators preferred the translation agency "A" translations significantly more. But even if the "A" translations were also better than "B" in from-Czech, we see it as very unlikely that the translatologist translations would be worse than all systems.

Source-based DA

(on document level)

SR+FD

English→Czech

Rank	Ave.	Ave. z	System
1	90.2	0.397	HUMAN-A
2-4	87.9	0.284	HUMAN-B
2-4	87.6	0.263	Facebook-AI
2-4	86.1	0.214	Online-W
5-7	83.0	0.122	eTranslation
5-6	82.1	0.047	CUNI-Transformer2018
6-8	79.2	-0.120	CUNI-DocTransformer
7-9	79.3	-0.154	CUNI-Marian-Baselines
8-10	77.8	-0.183	Online-B
9-10	74.6	-0.308	Online-A
11	76.2	-0.373	Online-Y
12	65.6	-0.674	Online-G

Five clusters

English→German

Rank	Ave.	Ave. z	System
1-10	83.3	0.209	Online-B
1-6	84.7	0.179	Online-W
1-10	86.6	0.109	WeChat-AI
1-6	87.6	0.077	Facebook-AI
3-11	86.8	0.008	UEdin
1-11	86.5	-0.014	P3AI
3-11	86.4	-0.031	HUMAN-A
3-11	86.1	-0.038	eTranslation
1-11	84.5	-0.063	HUMAN-C
10-12	84.5	-0.109	Online-Y
5-12	83.3	-0.131	happypoet
3-12	86.7	-0.134	NVIDIA-NeMo

Single cluster

English→Chinese

Rank	Ave.	Ave. z	System
1-8	74.9	0.205	HappyNewYear
1-5	82.5	0.186	HUMAN-B
1-7	81.2	0.139	Facebook-AI
1-5	80.0	0.105	HUMAN-A
3-9	75.5	0.045	Lan-Bridge-MT
2-11	81.0	0.019	bjtu_nmt
2-9	80.9	-0.012	SMU
7-12	75.3	-0.066	Borderline
4-12	75.7	-0.068	Machine_Translation
7-12	81.4	-0.074	capitalmarvel
8-12	79.3	-0.090	BUPT_rush
5-12	79.2	-0.105	NiuTrans

Single cluster

Contrastive, source-based DA

(segment level ignoring doc. context)

contr:SR−DC

English→Czech

Rank	Ave.	Ave. z	System
1-2	87.8	0.281	Facebook-AI
1-2	87.6	0.237	Online-W
3-5	85.6	0.091	CUNI-DocTransformer
3-6	84.9	0.067	CUNI-Transformer2018
4-7	84.3	0.026	HUMAN-A
3-6	84.1	-0.003	HUMAN-B
6-7	83.4	-0.057	eTranslation
8-9	82.7	-0.119	CUNI-Marian-Baselines
8-10	81.3	-0.219	Online-A
9-10	81.1	-0.238	Online-B
11-12	77.7	-0.489	Online-Y
11-12	75.8	-0.630	Online-G

Four clusters

English→German

Rank	Ave.	Ave. z	System
1–3	89.6	0.093	Facebook-AI
1–3	88.5	0.067	WeChat-AI
1–3	88.4	0.035	Online-W
4–9	87.2	-0.044	NVIDIA-NeMo
4–11	87.9	-0.058	HUMAN-C
4–10	86.7	-0.062	P3AI
4–9	86.5	-0.080	UEdin
4–10	87.1	-0.088	Online-B
4–10	86.9	-0.102	eTranslation
6–12	85.7	-0.190	happypoet
10–12	85.7	-0.192	Online-Y
10–12	85.8	-0.226	HUMAN-A

Two clusters

English→Chinese

Rank	Ave.	Ave. z	System
1–5	82.6	0.072	Borderline
1–5	82.3	0.071	bjtu_nmt
1–5	82.5	0.062	SMU
1–5	82.4	0.048	Facebook-AI
1–5	82.5	0.011	NiuTrans
6–11	82.0	-0.016	HappyNewYear
6–11	82.0	-0.016	Machine_Translation
6–10	82.0	-0.056	Lan-Bridge-MT
6–11	81.6	-0.094	BUPT_rush
6–11	81.2	-0.126	capitalmarvel
6–11	81.7	-0.149	HUMAN-A
12	79.3	-0.393	HUMAN-B

Three clusters

Table 11: Contrastive results of WMT21 News Translation Task for translation out-of-English. Systems ordered by DA score z-score; systems within a cluster are considered tied; lines indicate clusters according to Wilcoxon rank-sum test $p < 0.05$; rank ranges are based on the same test (for details, see Section 3.2.2); grayed entry indicates resources that fall outside the constraints provided. DA scores collected using a segment-level annotation interface, so context is not available to annotators.

Language Pair	Sys.	Assess.	Evaluators
Bengali→Hindi	9	4,461	2
Hindi→Bengali	9	4,512	2
Xhosa→Zulu	6	2,952	2
Zulu→Xhosa	5	2,502	1
Total	29	14,437	7

Table 12: Amount of data collected in the WMT21 manual evaluation campaign for evaluation Hindi to/from Bengali and Zulu to/from Xhosa

addition to all German→English submissions, four original transformer baselines with varying sizes and depths were trained and evaluated: GENIE-large-6-6 (transformer large with a 6-layer encoder and a 6-layer decoder), GENIE-base-6-6, GENIE-base-3-3, and GENIE-base-1-1.[19] These models were trained solely on the given training data without ensembling, backtranslation, or any other data augmentation method.

Similar to the official into-English evaluations, evaluations are done monolingually where Human-A is used as the reference. Each HIT contains 5 segments that are randomly shuffled, and no document context is considered during evaluations. Turkers are asked to decide whether they agree or disagree that the prediction adequately expresses the meaning of the reference. Turkers are given the following additional instructions: *a prediction is adequate if in the absence of the reference, the prediction perfectly conveys the meaning intended by the reference*. The user interface for annotating one candidate segment in the HIT is illustrated in Figure 7.

For quality control, we first selected Amazon Mechanical Turkers who had completed at least 5000 HITs with a 99+% approval rate and had a locale of US, GB, AU, or CA. They were then asked to carefully read the instructions and finish 10 sample questions created from WMT 2019 submissions and references. They were allowed to participate only when they correctly annotate 9 instances at least. In addition to this quality control at the entry point, we kept monitoring to detect spamming behavior. In particular, we randomly replaced 5% of the model predictions with sentences identical to the corresponding reference (Perfect Ref., similar to *good reference* in Section 3.2.1), and 5% of the model predictions with the reference from a different question (Wrong Ref.). We then randomly selected 800 examples from the test set to annotate. During annotation, we monitored how annotators labeled the Perfect Ref. and Wrong Ref. questions. Annotators that failed to both assign a high score to the Perfect Ref. and a low score to the Wrong Ref. questions were removed from the annotator pool, and all of their annotations were discarded. This qualification resulted in removing 5% of the participants. Since spammers invest little effort into completing each HIT, they can complete many more than other annotators (we found they would have completed up to 50% of the HITs in our preliminary experiments). Therefore, removing the 5% of participants that spammed annotations substantially improved the quality of our assessment.

In summary, there are several major differences from the setup used in the official evaluations:

- Turkers assess the adequacy by a five-category Likert scale, which is later converted to scalar values: *strongly agree* (1.0), *agree* (0.75), *neutral* (0.5), *disagree* (0.25), and *strongly disagree* (0.0).
- All 5 segments are randomly chosen for each HIT, and the document context is disregarded.
- For evaluating each system, we randomly sample 800 segments from the test set. The randomly selected instances are shared across all systems.
- To maximize the number of segments annotated for a given budget, each segment is annotated only once (*unilabeling*). Under a fixed annotation budget, unilabeling results are shown to be relatively stable compared to *multilabeling* (i.e., evaluating one segment by multiple annotators. See Section 5.1 of Khashabi et al., 2021).
- The overall scores are calculated by averaging raw numbers over the 800 segments. No standardization is applied.
- Different quality controls are applied as discussed above.

Table 14 shows results from the GENIE evaluation for German to English translation. There are systems that are ranked highly, both in the official and GENIE evaluations, such as Online-A and VolcTrans-AT. Conversely, happypoet and Manifold are given low scores consistently. Further, the

[19] The leaderboard is public at `https://leaderboard.allenai.org/genie-mt21/submissions/public`. All models and code to reproduce are available at `https://github.com/jungokasai/GENIE_wmt2021-de-en`.

Bengali→Hindi

Rank	Ave.	Ave. z	System
1–2	82.1	0.202	GTCOM
1–2	79.1	0.163	Online-B
3–5	77.5	0.080	TRANSSION
3–5	78.0	0.076	MS-EgDC
3–6	78.0	0.054	UEdin
4–8	76.1	−0.015	Online-Y
6–8	75.7	−0.080	HW-TSC
6–8	75.7	−0.107	Online-A
9	70.8	−0.373	Online-G

Hindi→Bengali

Rank	Ave.	Ave. z	System
1–4	95.0	0.245	HW-TSC
1–4	94.8	0.236	Online-A
1–4	94.5	0.233	GTCOM
1–4	94.6	0.214	UEdin
5–6	92.3	0.080	Online-Y
7	92.0	0.045	TRANSSION
6–7	91.3	0.029	Online-B
8	90.9	−0.008	MS-EgDC
9	73.5	−1.100	Online-G

Xhosa→Zulu

Rank	Ave.	Ave. z	System
1–3	68.4	0.331	HW-TSC
1–3	67.9	0.287	TRANSSION
1–3	63.7	0.240	GTCOM
4–5	61.5	0.144	MS-EgDC
4–5	62.6	0.107	FJDMATH
6	19.4	−1.135	Online-G

Zulu→Xhosa

Rank	Ave.	Ave. z	System
1	80.7	0.502	TRANSSION
2–3	74.3	0.310	HW-TSC
2–4	72.6	0.258	MS-EgDC
3–4	69.3	0.162	GTCOM
5	21.9	−1.253	Online-G

Table 13: Official results of WMT21 Translation Task for Hindi to/from Bengali and Zulu to/from Xhosa translation (Wikipedia data, SR−DC). Systems ordered by DA score z-score; systems within a cluster are considered tied; lines indicate clusters according to Wilcoxon rank-sum test $p < 0.05$; rank ranges are based on the same test (for details, see Section 3.2.2); grayed entry indicates resources that fall outside the constraints provided.

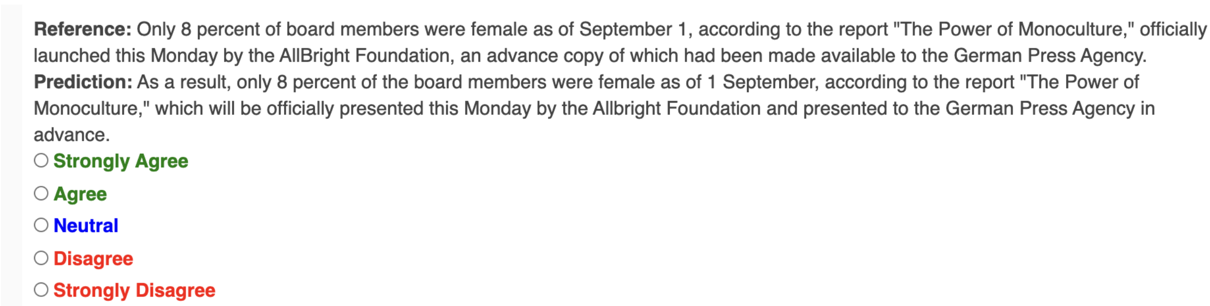


Figure 7: GENIE annotation interface for one segment.

transformer baselines are ranked in the expected order: large-6-6, base-6-6, base-3-3, followed by base-1-1. This confirms the validity of the evaluations. Nonetheless, we see some noticeable difference from the official ranking. In particular, HUMAN and the Watermelon systems are ranked high in contrast to the official evaluations. It is left to future work to analyze which parts of the crowdsourcing setup are contributing to the diverging system rankings; these analyses would help us improve our human evaluation method in the future.

4 Similar Language Translation

In this section we present the findings of the third SLT shared task organized at WMT 2021. The task follows the success of the two past SLT shared tasks organized at WMT 2019 and WMT 2020. SLT 2021 is motivated by the growing interest of the community in translating between similar languages, low-resource languages, dialects, and language varieties, and the challenges faced by state-of-the-art systems in these settings evidenced in recent studies (Hassani, 2017; Costa-jussà et al., 2018; Popović et al., 2020; Tapo et al., 2020).

The main goal of the task is to evaluate the performance of state-of-the-art MT systems on translating between closely-related language pairs of languages from the same language family. Past editions of the task (Barrault et al., 2019, 2020) featured language pairs such as Spanish - Portuguese, Czech - Polish, and Hindi - Nepali to name a few. This year's SLT features multiple pairs of similar languages from the Indo-Aryan and Romance family.

Finally, SLT 2021 also features a track including French and two similar low-resource Manding languages spoken in West Africa, namely Bambara and Maninka, where participants were pro-

GENIE German→English

Ave. Score	Lower	Upper	System
0.757	0.737	0.776	Watermelon
0.752	0.732	0.772	VolcTrans-AT
0.752	0.732	0.772	HUMAN
0.743	0.724	0.764	Online-B
0.742	0.721	0.760	Online-A
0.740	0.720	0.759	Facebook-AI
0.738	0.721	0.756	Online-W
0.738	0.717	0.757	Online-G
0.737	0.717	0.757	VolcTrans-GLAT
0.735	0.714	0.756	UF
0.734	0.713	0.754	HuaweiTSC
0.733	0.710	0.753	NVIDIA-NeMo
0.712	0.691	0.734	ICL
0.704	0.684	0.723	GENIE-large-6-6
0.704	0.684	0.722	P3AI
0.700	0.680	0.721	UEdin
0.692	0.670	0.712	SMU
0.690	0.669	0.711	GENIE-base-6-6
0.685	0.664	0.705	Manifold
0.676	0.655	0.696	Borderline
0.665	0.645	0.684	Online-Y
0.653	0.630	0.676	GENIE-base-3-3
0.643	0.620	0.667	happypoet
0.507	0.483	0.530	GENIE-base-1-1

Table 14: GENIE DE-EN results. Lower and upper bounds for 95% confidence intervals are calculated by bootstrapping (Koehn, 2004; Khashabi et al., 2021). Grayed entries indicate unconstrained settings.

vided with the opportunity to combine datasets of the two Manding languages taking advantage of their similarity. As in past editions of the task, translations at SLT 2021 are evaluated in all directions using three automatic evaluation metrics: BLEU, RIBES, and TER.

4.1 Data

Training We have made available a number of data sources for the SLT shared task. Some training datasets were used in the previous editions of the WMT News Translation shared task and were updated (News Commentary v16, Wiki Titles v3), while some corpora were newly introduced. We also used data collected from Opus (Tiedemann and Nygaard, 2004; Tiedemann, 2012)[20].

For the Spanish–Catalan language pair we used parallel corpora: Wiki Titles v3, ParaCrawl (Bañón et al., 2020), DOGC v2, and monolingual: Europarl v10 (Koehn, 2005), News Commentary v16, News Crawl, caWaC (Ljubešić and Toral, 2014) (see Table 15). Released corpora for the Spanish–Portuguese language pair included parallel datasets: Europarl v10 (Koehn, 2005), News Commentary v16, Wiki Titles v3, Tilde MODEL (Rozis and Skadiņš, 2017), JRC-Acquis (Steinberger et al., 2006), and monolingual corpora: Europarl v10 (Koehn, 2005), News Commentary v16, News Crawl (see Table 16). Moreover, corpora for the Romanian–Spanish language pair (see Table 17) and the Romanian–Portuguese language pair (see Table 18) contained parallel datasets: Europarl v8 (Koehn, 2005), Wiki Titles v3, Tilde MODEL (Rozis and Skadiņš, 2017), JRC-Acquis (Steinberger et al., 2006), and monolingual data: Europarl v10 (Koehn, 2005), News Commentary v16, News Crawl, Common Crawl.

The released parallel Tamil–Telugu dataset was collected from news (Siripragada et al., 2020), PMIndia (Haddow and Kirefu, 2020) and MKB (Man Ki Baat) datasets. All data were initially combined, tokenized using indic-nlp tokenizer (Kunchukuttan, 2020) and randomly shuffled. A subset of data extracted from the dataset are used for test and development set. The remaining data were considered as training set (cf. Table 21).

Finally, the parallel Bambara-French corpus is a part of the Bambara Reference Corpus [21].

Development and Test Data The development and test sets for Spanish–Catalan, Spanish–Portuguese, Romanian–Spanish and Romanian–Portuguese language pairs were created from a corpus provided by Pangeanic[22]. Catalan translations were provided by the Directorate-General for Language Policy at the Ministry of Culture, Government of Catalonia. Each dev and test dataset was cleaned, deduplicated and shuffled, resulting in 969 and 999 sentences in dev and test sets respectively.

4.2 Participants and Approaches

SEBAMAT SEBAMAT submitted their system for two language pairs, Spanish–Catalan and Spanish–Portuguese, in both directions. The SEBAMAT approach is based on the Marian NMT toolkit that leverages the Transformer architecture. The systems were trained using only the parallel corpora that were made available for the participants. For all the language pairs and directions, SEBAMAT submitted PRIMARY and CONTRASTIVE systems with different vocabulary sizes (40,000 and 85,000, respectively). Interestingly, in all the cases, the PRIMARY systems with a smaller vocabulary size performed better in terms of BLEU scores.

[20] http://opus.nlpl.eu/

[21] http://cormand.huma-num.fr/index.html

[22] https://www.pangeanic.com/

		Corpus	Sentences
Parallel	Spanish ↔ Catalan	Wiki Titles v3	476,475
	Spanish ↔ Catalan	ParaCrawl	6,870,183
	Spanish ↔ Catalan	DOGC v2	10,933,622
Monolingual	Spanish	Europarl v10	2,038,042
	Spanish	News Commentary v16	503,255
	Spanish	News Crawl 2007-2020	65,365,886
	Catalan	caWaC	24,745,986
Dev	Spanish ↔ Catalan		969
Test	Spanish ↔ Catalan		999

Table 15: Corpora for the Spanish ↔ Catalan language pair.

		Corpus	Sentences
Parallel	Spanish ↔ Portuguese	Europarl v10	1,801,845
	Spanish ↔ Portuguese	News Commentary v16	48,259
	Spanish ↔ Portuguese	Wiki Titles v3	649,833
	Spanish ↔ Portuguese	Tilde MODEL	13,464
	Spanish ↔ Portuguese	JRC-Acquis	1,650,126
Monolingual	Spanish	Europarl v10	2,038,042
	Spanish	News Commentary v16	503,255
	Spanish	News Crawl 2007-2020	65,365,886
	Portuguese	Europarl v10	2,016,635
	Portuguese	News Commentary v16	89,111
	Portuguese	News Crawl 2008-2020	10,900,924
Dev	Spanish ↔ Portuguese		969
Test	Spanish ↔ Portuguese		999

Table 16: Corpora for the Spanish ↔ Portuguese language pair.

		Corpus	Sentences
Parallel	Romanian ↔ Spanish	Europarl v8	387,653
	Romanian ↔ Spanish	Wiki Titles v3	253,770
	Romanian ↔ Spanish	Tilde MODEL	3,770
	Romanian ↔ Spanish	JRC-Acquis v2	451,849
Monolingual	Spanish	Europarl v10	2,038,042
	Spanish	News Commentary v16	503,255
	Spanish	News Crawl 2007-2020	65,365,886
	Romanian	Common Crawl	288,806,234
	Romanian	News Crawl 2015-2020	29,538,472
Dev	Romanian ↔ Spanish		969
Test	Romanian ↔ Spanish		999

Table 17: Corpora for the Romanian ↔ Spanish language pair.

T4T The T4T team participated in the SLT 2021 Romance languages track, submitting their system for Spanish ↔ Catalan and Spanish ↔ Portuguese. While their systems are built using out-of-the-box OpenNMT toolkit, the team developed custom cleaning scripts and an adhoc tokenizer. SentencePiece library was used for pre-processing and reducing the vocabulary size to 16,000 symbols.

UBC-NLP The UBC-NLP team submitted their Spanish ↔ Portuguese, Catalan → Spanish and French ↔ Bambara systems to the SLT 2021 task. Their systems are built using Transformers from the HuggingFace library. The UBC-NLP team experimented with tokenized (PRIMARY) and untokenized (CONTRASTIVE) systems and compared them with models developed by fine-tuning pre-trained models as well as models trained from

		Corpus	Sentences
Parallel	Romanian ↔ Portuguese	Europarl v8	381,404
	Romanian ↔ Portuguese	Wiki Titles v3	251,834
	Romanian ↔ Portuguese	Tilde MODEL	3,860
	Romanian ↔ Portuguese	JRC-Acquis v2	451,737
Monolingual	Portuguese	Europarl v10	2,016,635
	Portuguese	News Commentary v16	89,111
	Portuguese	News Crawl 2008-2020	10,900,924
	Romanian	Common Crawl	288,806,234
	Romanian	News Crawl 2015-2020	29,538,472
Dev	Romanian ↔ Portuguese		969
Test	Romanian ↔ Portuguese		999

Table 18: Corpora for the Romanian ↔ Portuguese language pair.

		Corpus	Sentences
Parallel	French ↔ Bambara	Dokotoro/Bible/SIL Dictionary Sentences/Corpus Référence de Bambara	9,939
Dev	French ↔ Bambara		5,972
Test	French ↔ Bambara		2,984

Table 19: Corpora for the French ↔ Bambara language pair.

		Corpus	Sentences
Parallel	French ↔ Maninka	3000 training sentences/Constitution of Guinea	3,243
Dev	French ↔ Maninka		540
Test	French ↔ Maninka		270

Table 20: Corpora for the French ↔ Maninka language pair.

		Corpus	Sentences
Parallel	Tamil ↔ Telugu	MKB	3,100
	Tamil ↔ Telugu	News	11,038
	Tamil ↔ Telugu	PM India	26,009
Dev	Tamil ↔ Telugu		1,261
Test	Tamil ↔ Telugu		1,735

Table 21: Corpora for the Tamil ↔ Telugu language pair.

scratch. The pre-trained models were developed using Marian NMT by Helsinki-NLP on HuggingFace.

A3-108 The A3-108 team submitted 3 systems (one PRIMARY and two CONTRASTIVEs) based on statistical machine translation for Tamil ↔ Telugu language pair. The team explores various tokenization schemes for their submissions. Their PRIMARY run achieved top rank in Telugu → Tamil and ranked 3^{rd} in Tamil → Telugu translation task.

oneNLP oneNLP team participation on Tamil ↔ Telugu system is based on transformer based NMT. The team explored different subword configurations, script conversion and single model training for both directions. Their primary submission achieved 2.05 BLEU for Tamil → Telugu and 5.03 for Telugu → Tamil.

CNLP-NITS The team submitted their run for Tamil ↔ Telugu similar language translation task. The CNLP-NITS system used pre-train word embeddings from monolingual data and applied in transformer based neural machine translation. The model achieved BLEU score 4.05 for both Tamil → Telugu and Telugu → Tamil.

NITK-UOH NITK-UoH's submission system is based on vanilla Transformer model initialized with MultiBPEmb – a collection of multilingual subword segmentation based pretrained embeddings. NITK-UoH performs top in Tamil → Telugu translation task.

4.3 Results

Similarly to the previous edition of the SLT shared task, participants could submit systems for the Spanish–Catalan and Spanish–Portuguese language pairs (in both directions). The best systems for Spanish-to-Portuguese (see Table 25) achieved over 40 BLEU and around 85 RIBES. While in the opposite direction (Portuguese-to-Spanish) the best performing system reached 47.71 of BLEU (see Table 24). As the Spanish–Catalan dev and test sets were aligned with Spanish–Portuguese ones, we noticed that the best results for the Spanish–Catalan language pair are in general much better than for Spanish–Portuguese. For Spanish-to-Catalan the best system attained over 79 BLEU and below 15 TER (see Table 27). However, its RIBES score (95.76) was lower than the runner-up system's (96.24). In the case of Catalan-to-Spanish, the best system scored over 82 BLEU and less than 11 TER (see Table 26). As there were no submissions for Romanian–Spanish and Romanian–Portuguese, we do not provide any evaluations for these language pairs.

4.4 Summary

This section presented the results and findings of the third edition of the SLT shared task at WMT. The third iteration of this competition featured data from multiple language pairs from three different language families: Dravidian, Manding, and Romance languages. We evaluated the systems translating in both directions of the language pair using three automatic metrics: BLEU, RIBES, and TER. Most teams this year participated in the Dravidian language pairs. Following a trend observed in the past editions of the task, we observed that the performance varies widely between language pairs and domains.

5 Triangular MT

This section presents an overview of the Triangular MT shared task. Given a low-resource language pair (X/Y), the bulk of previous MT work has pursued one of two strategies.

- Direct: Collect parallel X/Y data from the web, and train an X-to-Y translator , OR
- Pivot (Utiyama and Isahara, 2007; Wu and Wang, 2009): Collect parallel X/English and Y/English data (often much larger than X/Y data), train two translators (X-to-English + English-to-Y), and pipeline them to form an X-to-Y translator

However, there are many other possible strategies for combining such resources. These may involve, for example, ensemble methods, multi-source training methods, multi-target training methods, or novel data augmentation methods. For eg. (Zoph et al., 2016; Dholakia and Sarkar, 2014; Kim et al., 2019).

5.1 The Task

The goals of this shared task is to promote:

- translation between non-English languages,
- optimally mixing direct and indirect parallel resources, and
- exploiting noisy, parallel web corpora

The task is Russian-to-Chinese machine translation. We provided parallel corpora to the participating teams. We evaluate system translations on a (secret) mixed-genre test set, drawn from the web and curated for high quality segment pairs. After receiving test data, participants had one week to submit translations. After all submissions are received, we posted a populated leaderboard that will continue to receive post-evaluation submissions.[23] The evaluation metric for the shared task is 4-gram character Bleu. The script to be used for Bleu computation is Moses `multi-bleu-detok.perl`. Instructions to run the script were released as part of the shared task.[24] The participants indicated their intent to participate via registration on the Codalab website for the shared task[25] and obtained the instructions and links to various resources.

5.2 Training Data

We provided three parallel corpora:

- Chinese/Russian: crawled from the web and aligned at the segment level, and combined with different public resources.

[23] https://competitions.codalab.org/competitions/30446#results

[24] https://github.com/didi/wmt2021_triangular_mt/tree/master/eval

[25] https://competitions.codalab.org/competitions/30446#participate

Team Name	System Type	BLEU ↑	RIBES ↑	TER ↓
NITK-UOH	PRIMARY	6.09	17.03	-
A3-108	CONTRASTIVE1	5.54	40.58	98.082
A3-108	PRIMARY	5.23	42.37	98.662
CNLP-NITS	PRIMARY	4.05	24.80	97.241
oneNLP	CONTRASTIVE2	3.67	22.28	99.122
oneNLP	CONTRASTIVE	3.57	23.54	99.034
A3-108	CONTRASTIVE2	3.32	34.42	-
oneNLP	PRIMARY	2.05	21.68	-
NITK-UOH	CONTRASTIVE	0.00	0.03	-

Table 22: Evaluation results for Tamil to Telugu.

Team Name	System Type	BLEU ↑	RIBES ↑	TER ↓
A3-108	PRIMARY	8.37	43.55	95.884
A3-108	CONTRASTIVE1	7.89	46.24	95.627
A3-108	CONTRASTIVE2	7.43	42.54	94.964
NITK-UOH	PRIMARY	6.55	19.61	98.356
oneNLP	PRIMARY	5.03	23.98	97.551
CNLP-NITS	PRIMARY	4.05	24.80	97.241
oneNLP	CONTRASTIVE	3.63	27.05	97.534
oneNLP	CONTRASTIVE2	3.61	26.12	96.772
NITK-UOH	CONTRASTIVE	0.04	1.00	-

Table 23: Evaluation results for Telugu to Tamil.

Team Name	System Type	BLEU ↑	RIBES ↑	TER ↓
UBC-NLP	PRIMARY	47.71	87.11	39.213
SEBAMAT	PRIMARY	46.51	86.31	41.235
T4T	PRIMARY	46.29	87.04	40.181
UBC-NLP	CONTRASTIVE	43.86	85.10	43.801
SEBAMAT	CONTRASTIVE	43.12	84.99	45.068

Table 24: Evaluation results for Portuguese to Spanish.

Team Name	System Type	BLEU ↑	RIBES ↑	TER ↓
T4T	PRIMARY	40.74	85.69	43.343
SEBAMAT	PRIMARY	40.35	84.99	45.258
SEBAMAT	CONTRASTIVE	38.90	83.89	47.044
UBC-NLP	PRIMARY	38.10	85.35	46.556
UBC-NLP	CONTRASTIVE	35.61	82.48	52.612

Table 25: Evaluation results for Spanish to Portuguese.

Team Name	System Type	BLEU ↑	RIBES ↑	TER ↓
UBC-NLP	PRIMARY	82.79	96.98	10.918
SEBAMAT	PRIMARY	78.65	94.76	15.805
T4T	PRIMARY	77.93	96.04	16.502
UBC-NLP	CONTRASTIVE	76.8	95.19	15.421
SEBAMAT	CONTRASTIVE	76.78	94.46	17.067

Table 26: Evaluation results for Catalan to Spanish.

Team Name	System Type	BLEU ↑	RIBES ↑	TER ↓
SEBAMAT	PRIMARY	79.69	95.76	14.632
T4T	PRIMARY	78.60	96.24	16.133
SEBAMAT	CONTRASTIVE	77.32	95.35	16.744

Table 27: Evaluation results for Spanish to Catalan.

Team Name	System Type	BLEU ↑	RIBES ↑	TER ↓
UBC-NLP	PRIMARY	1.32	24.79	97.899

Table 28: Evaluation results for French to Bambara.

- Chinese/English: combining several public resources.
- Russian/English: combining several public resources.

The details of the training resources provided are shown in Table 30. The provenance of the collected parallel data is as follows. We used a parallel data harvesting pipeline developed at DiDi (Zhang et al., 2020) to harvest Russian/Chinese parallel data on the Internet. We downloaded parallel datasets available from Opus (Tiedemann, 2009) for all the three language pairs - Russian/Chinese, Russian/English and English/Chinese. Since united nations data and subtitles data (Ru/En) are very large sources of parallel data, we report statistics on these two types of Opus parallel sources. In addition to Opus, we also curate parallel data from Wikimatrix (Schwenk et al., 2019) in all three language pairs and social media parallel data - Weibo and Twitter (Ling et al., 2013). We also release the provenance of each parallel segment, in case teams want to use this information to filter noisy data sources.

5.3 Creating the Test Dataset

We spent a considerable amount of time to curate high quality, parallel data online to be used as development and evaluation datasets. This was a completely manual process undertaken by a native speaker of Russian who consulted with a native Chinese speaker from our team to ensure good quality translations (that does not contain tell-tale signs of automatic translation). Our workflow entailed finding websites and large chunks of parallel text, not necessarily from the same pages. The sources selected were also hard to be harvested from a parallel data pipeline due to their difference in URL structure. The sources selected were from a diverse range of non-traditional sources, and have a balance of different types of documents. The topics would be famous works of literature, or tourism related news stories, and so on. We copied large chunks of text from such sources and manually aligned the paragraphs, followed by manual sentence alignment, each done manually to ensure top quality parallel segments. This was followed by a final filtering step to remove sentences and entire sources which had a significant overlap with training and development data. The details of the development and test datasets are shown in Tables 31 and 32.

5.4 Baselines and Final Results

We released a baseline system[26] as part of the shared task. This is based on the Google Tensor2tensor[27] toolkit to train a Transformer-based NMT system. We also provided the baseline bleu score on the development dataset ahead of the evaluation phase. We had 2 simple baselines - (1) Direct - Transformer model trained on the entire Russian/Chinese parallel dataset and decoded with $\alpha = 1.0$ and $beam\_size$=4. (2) Pivot model - 2 MT systems - Russian-to-English and English-to-Chinese - each trained with the corresponding parallel data. Both the Russian-to-English and the English-to-Chinese systems were decoded with $alpha$=1.0 and $beam\_size$=4. The baseline results on the development dataset as shown in Table 33.

We had a total of six teams submitting their system outputs on the test dataset. The evaluation metric was 4-gram character bleu score. The final evaluation results are shown in Table 34.

[26] https://github.com/didi/wmt2021_triangular_mt/

[27] https://github.com/tensorflow/tensor2tensor

Team Name	System Type	BLEU ↑	RIBES ↑	TER ↓
UBC-NLP	PRIMARY	3.62	36.17	-

Table 29: Evaluation results for Bambara to French.

Russian/Chinese parallel data	**Segment pairs**	**Characters (Chinese side)**
DiDi parallel data harvesting pipeline	5,403,157	82,552,922
Opus (no UN) + Weibo + Wikimatrix	430,302	20,954,541
Opus (UN)	27,551,996	1,362,478,536
Total	*33,385,455*	*1,465,985,999*
Russian/English parallel data	**Segment pairs**	**Words (Russian side)**
Opus (no UN, no subtitles) + Twitter + Wikimatrix	6,340,245	97,537,275
Opus (UN, subtitles)	62,811,986	909,476,736
Total	*69,152,231*	*1,007,014,011*
English/Chinese parallel data	**Segment pairs**	**Characters (Chinese side)**
Opus (no UN) + Twitter + Weibo + Wikimatrix	1,435,132	69,894,886
Opus (UN)	27,089,931	1,333,732,823
Total	*28,525,063*	*1,403,627,709*

Table 30: Triangular MT: Training data statistics

5.5 Overview of the Submitted Systems

Five out of the six participating systems submitted system description papers. In this section we briefly discuss the outline of these systems. For more details please refer to the proceedings.

- istic-team-2021 (Guo et al., 2021) The team's system is based on the Transformer architecture. They used several corpus pre-processing steps such as special symbol filtering and filtering based on segment length. In addition, they used context-based system combination - which is a multi-encoder to encode source sentence and contextual information from the machine translation results on the source sentence. They tried with both a direct and pipeline-based pivot system and report that the latter outperforms the former.

- HW_TSC (Li et al., 2021a) Huawei's submission used a multilingual model which is a single neural machine translation model to translate among multiple languages. Upon adding more parallel data, they report an increase in bleu score of upto 2 points using the multilingual model compared to the baseline model. In addition they used several data pre-processing techniques to denoise the training data and data augmentation techniques such as back-translation to improve overall system performance.

- Papago (Park et al., 2021) Naver's system reports that they get better performance by treating this as a bilingual machine translation task rather than as a multilingual translation task, based on their early experiments. They use the transformer model with extensive data pre-processing, filtering and data augmentation. To augment the direct bilingual data they synthetically generate bilingual sentence pairs using monlingual Chinese back-translated to Russian and the 2 sets of indirect parallel dataset provided.

- DUT-MT (Liu et al., 2021a) This team experimented with 2 different multilingual training models called mBART and mRASP, both of them based on underlying Transformer architecture. They report boosted performance especially on rare words when using mRASP. In addition, they also carry out data preprocessing and filtering to improve system performance.

- CFILT-IITB (Mhaskar and Bhattacharyya, 2021) CFLIT-IITB team's system used a pivot-based transfer learning technique. In this technique they have 2 encoder-decoder models, source-pivot (Russian-to-English) and pivot-target (English-to-Chinese), each of them trained on the respective training datasets. They use the encoder of the former and the decoder of the latter to initialize a third encoder-decoder for the actual task of Russian-to-Chinese translation. They fine tune this decoder using the given parallel data for Russian/Chinese. They report this system has a better performance compared to either a direct or pivot-based cascaded system. They do not experiment much with data pre-processing and filtering.

Source	Genre	Parallel segments
Anna Karenina, dialog	Literature	98
Art Academy	Biography	67
Isaac Babel interview	Literature	104
Master and Margarita	Literature	106
MPMCMS	International news	71
Potato system	International news	97
Visit Amur	Tourism	250
Chinese Embassy in Russia	International news	172
Total	-	*965*

Table 31: Triangular MT: Development dataset details

Source	Genre	Parallel segments
Aeroflot	Tourism	99
Isaac Babel - salt	Literature	47
A Day Without Lies	Literature	200
Everything is Normal, Everything is Fine	Literature	98
Hujiang	Language Learning	236
Kazinform	Tourism	21
Lotos shopping centre	Tourism	17
Alexandra Marinina novel	Literature	55
Private Museum Catalog	Tourism	196
Solzhenitsyn Nobel speech	Literature	240
Russia Beyond	Biography	329
Shenyang consulate	International news	113
War and Peace	Literature	3
Russian Embassy in China	Tourism, International News	97
Total	-	*1751*

Table 32: Triangular MT: Test dataset details

5.6 Conclusion

The triangular machine translation shared task set out to explore various modeling possibilities when building a machine translation system for a non-English language pair. We received enthusiastic participation from the participants. Almost all of them performed data filtering and preprocessing to denoise the training datasets and that seemed to substantially help improve system performance. The transformer model and its variants were used in all the system submissions confirming Transformer's ubiquitous acceptance as the model of choice for building machine translation systems. Many teams explored model ensembling and model averaging in addition to model reranking strategies. Several teams explored back-translation as an effective data-augmentation strategy. There was a wide variety of modeling architectures experimented by the participants. Almost everyone used all the parallel datasets provided underlining the importance of using parallel data in all directions to build a better machine translation system. Overall we are happy that the shared task provided a platform to the participants to experiment with different modeling strategies. We hope practitioners will find these techniques useful when working on machine translation between non-English language pairs.

6 Multilingual Low-Resource Translation for Indo-European Languages Task

Massively multilingual machine translation has shown impressive results, including zero and few-shot translation of low-resource languages. However, these models are often evaluated from or into English, where the most data is available, and one assumes that the models would generalise to other language pairs and low-resource languages. This shared task focuses explicitly on checking this assumption and aims to explore multilingual archi-

System	BLEU
Google Translate API	33.04
BASELINE-DIRECT	20.24
BASELINE-PIVOT	19.33

Table 33: Triangular MT: Baseline results on the development dataset

	Team name	BLEU
	Google Translate API	30.2
Team 1	HW_TSC	27.7
Team 2	Papago	26.8
Team 3	DUT-MT	21.7
Team 4	istic-team-2021	19.2
Team 5	CFILT-IITB	18.8
-	BASELINE-PIVOT	17.9
-	BASELINE-DIRECT	17.0
Team 6	mcairt	16.6

Table 34: Triangular MT: Results on the test dataset

tectures for languages in a same family and evaluate only low-resource pairs even if using the high-resourced pairs in the same language family is not forbidden. We work in the cultural heritage domain, where we can consider full documents, and in two Indo-European language families: North-Germanic and Romance. With these goals in mind (multilinguality, specific domain and document-level translation) we define two tasks, one per family:

Task 1. Europeana thesis abstracts and descriptions. North-Germanic languages: from/to Icelandic (is), Norwegian Bokmål (nb) and Swedish (sv). Danish (da), German (de) and English (en) data is allowed for training but translation quality is not evaluated.

Task 2. Wikipedia cultural heritage articles. Romance languages: from Catalan (ca) to Occitan (oc), Romanian (ro) and Italian (it). Spanish (es), French (fr) and Portuguese (pt) data (+ English) is allowed for training but translation quality is not evaluated.

6.1 Data and Resources

6.1.1 Training Corpora

One of the purposes of the shared task is to obtain state-of-the-art systems for the language pairs in the domain involved. In principle, this would imply an unconstrained data setting but, we also want to be able to compare systems and architectures among themselves. For this, we constrain the amount of parallel and monolingual corpora to be used but we allow pretrained open-source systems which might use more data than allowed for the languages considered. All the sources listed below apply to the following languages (except for pre-trained models): Icelandic, Norwegian Bokmål, Swedish, Danish, German and English (Task 1); and Catalan, Italian, Occitan, Romanian, Spanish, French, Portuguese and English (Task 2).

- Corpora available at ELRC.[28] This data includes Paracrawl and Global voices.
- Europarl, JW300, WikiMatrix, MultiCCAligned, OPUS-100, Books, the Bible and TED talks.
- Common Crawl, Wikipedia and Wikidata dumps.
- Wordnets with open license, BabelNet.
- (Multiligual) pre-trained embeddings or other models that can be found freely available online (Hugging Face).
- Additional resources in Section 6.1.2 (multilingual lexicons).

6.1.2 Additional Resources

Given the importance of named entities in the cultural heritage domain, we provide participants with parallel/multilingual lexicons from Wikidata, Wikipedia titles and Wiktionary. The figures for each source are summarised in Table 35.

[28]https://elrc-share.eu/repository/search/

	Wikidata		Wikipedia		Wiktionary
	all	cleaner	all	cleaner	all
is2nb/nb2is	1,141,891	–	–	–	3,304/6,552
is2sv/sv2is	1,149,894	–	–	–	15,369/17,321
nb2sv/sv2nb	2,648,493	–	–	–	9,390/7,124
is-nb-sv	1,139,493	23,574	–	–	–
ca2it/it2ca	3,072,380	–	323,055	–	18,684/19,050
ca2oc/oc2ca	1,300,979	–	71,854	–	3,999/3,538
ca2ro/ro2ca	1,608,860	–	123,215	–	11,990/12,034
it2oc/oc2it	1,285,771	–	75,542	–	7,225/6,332
it2ro/ro2it	4,547,649	–	215,296	–	20,898/20,442
ro2oc/oc2ro	1,230,752	–	64,800	–	4,586/4,350
ca-it-ro	1,579,345	123,543	117,543	97,484	–

Table 35: Number of entries of the parallel/multilingual lexicons extracted from Wikidata, Wikipedia titles and Wiktionary for the multilingual low-resource translation task.

	Validation				Test			
	Docs.	Sents.	Src toks.	Tgt toks.	Docs.	Sents.	Src toks.	Tgt toks.
is2nb	26	467	6,096	6,932	24	563	8,256	9,301
is2sv	26	467	6,096	6,611	24	563	8,256	8,819
nb2is	19	502	7,673	7,495	16	540	9,218	8,867
nb2sv	19	502	7,673	7,499	16	540	9,218	8,804
sv2is	43	516	9,097	9,524	44	547	9,642	9,733
sv2nb	43	516	9,097	9,232	44	547	9,642	9,787
ca2it	41	1,269	30,363	29,725	42	1,743	38,868	37,649
ca2oc	41	1,269	30,363	30,184	42	1,743	38,868	38,662
ca2ro	41	1,269	30,363	29,842	42	1,743	38,868	37,379

Table 36: Statistics on the validation and test sets of the multilingual low-resource translation task. Source (Src) are original documents and target (Tgt) are human translations.

Wikidata. We extract aligned lexicons from the wikidata-20210301-all.json dump and provide two versions. The complete ("all") version includes all the entries, including duplicates. The "cleaner" version excludes duplicates, most of the terms that are equal in all the languages, terminology related to Wikimedia and a naïve cleaning on terms including years, parenthesis, and others.

Wikipedia titles. We extract aligned titles for the languages in Task 2 from the May 2020 Wikipedia dumps using the Wikitailor Toolkit[29] (Barrón-Cedeño et al., 2015; España-Bonet et al., 2020). We also provide two versions: the complete version ("all") includes all the entries. The "cleaner" version results from a naïve cleaning on titles including years, dates, parenthesis, and others.

Wiktionary. Each Wiktionary entry contains a word, its translation into several languages and its part of speech. We extract bilingual entries from April 2021 dumps for adjectives, adverbs, nouns and verbs from the Icelandic, Swedish, English and German Wiktionaries (Task 1) and from the Catalan and English ones (Task 2). The part of speech is kept in the dictionaries. Since the xlm dump contains the information in a text element with different structure for different dictionaries, we provide the extraction scripts for reproducibility.[30]

6.1.3 Validation and Test Sets

The documents used for constructing the validation and test sets are obtained from the Europeana collection (Task 1) and Wikipedia (Task 2).

Europeana kindly provided us with thesis abstracts, descriptions of archaeological sites and bibliographic entries for Icelandic, Norwegian Bokmål and Swedish. These monolingual documents are available at the Europeana portal but no intra-family parallel data exists and even the monolingual extraction is not straightforward for two main reasons: (i) collections with pan-Scandinavian labels and descriptions are uncommon, and (ii) language attributes in general are uncommon. For documents tagged as Norwe-

[29] github.com/cristinae/WikiTailor

[30] github.com/LeHarter/Extracting-translations-from-wiktionary

gian there is no distinction between Bokmål and Nynorsk, so texts where classified according to simple heuristics based on lexicons.

The original Europeana crawl obtained 1,192 documents (150,080 tokens) for Icelandic, 2,000 documents (166,303 tokens) for Norwegian Bokmål and 2,046 bilingual documents in English and Swedish with 443,111 tokens for Swedish. From these sets, we eliminate very similar documents (specially for Icelandic) and split documents at sentence level manually; we selected documents to collect around 1,000 sentences per language. Documents are finally divided evenly to build a validation set and a test set (Table 36).

The Wikipedia sets were built from articles in the Catalan edition. We selected original articles in Catalan that have no comparable article in any other language and that cover the cultural heritage domain (food, locations, sport, literature, traditions, people and animals). We selected 83 articles which were sentence-split manually to gather 3,013 sentences and 69,231 tokens. Similarly to the North-Germanic family, documents are divided evenly to build a validation set and a test set (Table 36). In this case, we also marked some entities in the source test documents (dates and locations) for further analysis in the manual evaluation (see Section 6.4).

Validation and test sets were sent to professional translators. A first translation was done by a native professional translator and afterwards there was a quality evaluation check by a second native professional translator. For the North-Germanic languages, we translated the source texts in Icelandic, Norwegian Bokmål and Swedish into the other two languages. For the Romance languages, we translated the source texts in Catalan into Italian, Romanian and Occitan. Translators were asked to keep the same sentence division as in the source and no indications were given on the translation of named entities.

6.2 Baselines and Submitted Systems

Nine different teams downloaded the validation data set but only five of them participated: BSC, CUNI, EdinSaar, Tencent and UBCNLP. We allowed two submissions per group and task, a primary (P) and a contrastive (C) system. With these constraints, we received four submissions for Task 1 and seven submissions for Task 2. We also prepared two baseline systems for comparison purposes.

6.2.1 M2M-100 (baseline)

We use M2M-100 without any modification, a multilingual model trained on a data set with 7.5 billion sentences for 100 languages including all the languages in our task (Fan et al., 2020). The sequence-to-sequence system is trained with parallel data enriched with backtranslations. We use the model with 1.2 B parameters available at the Hugging Face site.[31]

6.2.2 mT5-devFinetuned (baseline)

mT5 is a sequence-to-sequence model pretrained on a masked language modeling span-corruption objective with 8.5 billion monolingual sentences from 101 languages (Xue et al., 2021). As baseline, we use the model with 580 M parameters from Hugging Face. We finetune mT5-base only with the multilingual validation sets for each task described in Section 6.1.3. For Task 1, that involves 5,500 sentences, where we use the parallel sentences L_1–$\mathrm{L}_{2\mathrm{dev}}$ in both directions $\mathrm{L}_1 2\mathrm{L}_2$ and $\mathrm{L}_2 2\mathrm{L}_1$ (that is, we use is2nb$_{\mathrm{dev}}$ sentences as is2nb and nb2is, and nb2is$_{\mathrm{dev}}$ sentences as nb2is and is2nb because is2nb$_{\mathrm{dev}}$ and nb2is$_{\mathrm{dev}}$ are different; the same for the other pairs). We prepend one of the extra_id tokens in mT5 vocabulary to the source sentences to indicate the language of the target sentences. The remaining 440 sentences are used for validation. We repeat the process for Task 2, but in this case the training is multilingual but not bidirectional, so sentences are only used in one direction with a total of 3,600 sentences (1,200 ca2it, 1,200 ca2ro and 1,200 ca2oc) for finetuning and 207 for validation.

6.2.3 BSC (Kharitonova et al., 2021) – Task 2

BSC submission is a multilingual semi-supervised machine translation model. It is based on a pretrained language model, XLM-RoBERTa, that is later finetuned with parallel data obtained mostly from OPUS (5.1 M sentences). XLM-RoBERTa is only used to initialize the encoder while the shallow decoder is randomly initialised.

6.2.4 CUNI (Jon et al., 2021) – Task 2

Multilingual supervised machine translation model (primary) enriched with backtranslated data (contrastive). The multilingual systems

[31] https://huggingface.co/facebook/m2m100_1.2B

	Average Ranking	BLEU	TER	chrF	COMET	BertScore
M2M-100 (baseline)	1.0±0.0	31.5	0.54	0.55	0.399	0.862
EdinSaar-Contrastive	2.2±0.4	27.1	0.57	0.54	0.283	0.856
EdinSaar-Primary	2.8±0.4	27.5	0.58	0.52	0.276	0.849
UBCNLP-Primary	4.0±0.0	24.9	0.60	0.50	0.076	0.847
UBCNLP-Contrastive	5.0±0.0	24.0	0.61	0.49	-0.068	0.837
mT5-devFinetuned (baseline)	6.0±0.0	18.5	0.78	0.42	-0.102	0.810

Table 37: Official ranking according to the automatic metric average for the multilingual low-resource translation task of Europeana documents for North-Germanic languages (Task 1).

	Average Ranking	BLEU	TER	chrF	COMET	BertScore
CUNI-Primary	1.2±0.4	50.1	0.401	0.694	0.566	0.901
CUNI-Contrastive	1.6±0.5	49.5	0.404	0.693	0.569	0.901
TenTrans-Contrastive	3.0±0.0	43.5	0.460	0.670	0.444	0.894
TenTrans-Primary	3.8±0.4	43.3	0.462	0.668	0.442	0.894
BSC-Primary	5.0±0.7	41.3	0.402	0.647	0.363	0.884
M2M-100 (baseline)	5.8±0.4	40.0	0.478	0.634	0.414	0.878
UBCNLP-Primary	7.2±0.4	35.4	0.528	0.588	0.007	0.854
mT5-devFinetuned (baseline)	8.0±0.7	29.3	0.592	0.553	0.059	0.850
UBCNLP-Contrastive	8.6±0.5	28.5	0.591	0.529	-0.374	0.825

Table 38: Official ranking according to the automatic metric average for the multilingual low-resource translation task of Wikipedia articles in the cultural heritage domain for Romance languages (Task 2).

use 41 M original parallel sentences including all language pairs in the task plus French and English. Besides leveraging multilingual training data, various subword granularities are explored and phonemic representation of texts are added via multi-task learning. For Catalan–Occitan, character-level rescoring on the translations n-best lists is applied and Apertium is used for backtranslations when included.

6.2.5 EdinSaar (Tchistiakova et al., 2021) – Task 1

Semi-supervised systems with multilingual pretraining, backtranslation, finetuning and checkpoint ensembling. The primary system is a semi-supervised machine translation model. mT5 is finetuned with 1.2 M parallel sentences in the languages of the task plus Danish, German and English. The contrastive system is a transformer base architecture trained with 422 M parallel sentence pairs in all 30 language directions (including Danish, German and English) and finetuned only with pairs with the languages of the task as target language.

6.2.6 TenTrans (Yang et al., 2021) – Task 2

TenTrans submissions are semi-supervised multilingual systems based on a transformer base architecture. The basic system is an 8-to-4 multilingual model with Catalan–Italian–Romanian–Occitan as the target side and the inclusion of the high resource languages Spanish, French, Portuguese and English on the source side. In-domain finetuning is done with data selected using a domain classifier trained with multilingual BERT. Knowledge transfer is achieved with knowledge distillation of the M2M 1.2B model previously finetuned on the languages of the task. The primary submission is an ensemble between the in-domain multilingual and the distilled M2M. The contrastive submission adds a multilingual base model enriched with backtranslations to the ensemble and pivot-based methods to augment the training corpus.

6.2.7 UBCNLP (Chen and Abdul-Mageed, 2021) – Task 1, Task 2

Supervised bilingual systems based on a transformer base architecture where the Helsinki-NLP pretrained models available at the Hugging Face site are finetuned to the languages of the shared task. The primary submission finetunes the

	sv2nb					is2nb				
	BLEU	TER	chrF	COMET	BertSc	BLEU	TER	chrF	COMET	BertSc
M2M-100	**56.8**	**0.29**	**0.77**	**1.048**	**0.935**	19.3	0.67	0.42	-0.133	0.825
mT5-dFT	36.3	0.46	0.63	0.716	0.891	**22.3**	**0.64**	**0.47**	**0.120**	**0.853**
EdinSaar-C	48.2	0.35	0.73	0.980	0.923	13.0	0.71	0.41	-0.250	0.820
EdinSaar-P	45.4	0.38	0.70	0.919	0.912	16.3	0.72	0.39	-0.287	0.812
UBCNLP-C	51.8	0.33	0.74	0.996	0.931	9.5	0.77	0.33	-0.827	0.778
UBCNLP-P	49.8	0.35	0.73	0.952	0.927	12.8	0.74	0.36	-0.628	0.799

	nb2is					sv2is				
	BLEU	TER	chrF	COMET	BertSc	BLEU	TER	chrF	COMET	BertSc
M2M-100	**21.5**	**0.64**	**0.47**	**0.259**	**0.833**	19.0	0.66	0.48	0.501	0.832
mT5-dFT	3.6	1.26	0.21	-0.986	0.705	9.4	0.82	0.35	-0.138	0.777
EdinSaar-C	18.3	0.66	0.46	0.155	0.829	20.2	0.65	0.50	0.469	0.836
EdinSaar-P	19.5	0.65	0.46	0.258	0.829	**22.4**	**0.64**	**0.51**	**0.509**	0.836
UBCNLP-C	7.8	0.78	0.32	-0.924	0.771	20.5	0.66	0.49	0.348	**0.838**
UBCNLP-P	15.7	0.68	0.43	-0.074	0.822	14.8	0.71	0.45	0.144	0.825

	nb2sv					is2sv				
	BLEU	TER	chrF	COMET	BertSc	BLEU	TER	chrF	COMET	BertSc
M2M-100	**50.9**	**0.34**	**0.72**	**0.826**	**0.921**	**21.2**	**0.63**	0.45	-0.110	0.826
mT5-dFT	18.6	0.82	0.40	-0.368	0.790	21.1	0.69	**0.46**	**0.047**	**0.844**
EdinSaar-C	45.4	0.37	0.69	0.690	0.911	17.3	0.66	0.42	-0.348	0.815
EdinSaar-P	42.9	0.40	0.65	0.615	0.898	18.8	0.68	0.41	-0.357	0.805
UBCNLP-C	36.8	0.43	0.63	0.422	0.893	17.6	0.69	0.40	-0.425	0.810
UBCNLP-P	42.7	0.39	0.67	0.636	0.906	14.0	0.70	0.38	-0.572	0.804

Table 39: Automatic evaluation per language pair in the North-Germanic family of the multilingual low-resource translation task (Task 1). Best scores boldfaced. Notice that the final ranking is done per family and not per language pair as shown in Table 37.

Catalan–Spanish Helsinki-NLP model with WikiMatrix data (1.1 M sentences for ca-it, 139 k for ca-oc and 490 k for ca-ro). The same data is used to finetune the Catalan–English Helsinki-NLP model in the contrastive submission.

6.3 Automatic Evaluation

Recently, automatic metrics based on contextual embeddings have been shown to correlate better than string matching ones with human judgments (Kocmi et al., 2021). COMET was shown to be the best performing metric for languages with Latin script and chrF the best performing string-based method. Still, BLEU is used as *de facto* metric in most papers. As we cannot perform human evaluation for the 9 language pairs involved in this shared task, for the official ranking we use a combination of several metrics including the ones just mentioned plus BertScore as representative of contextual embedding-based metrics and TER as representative of plain string methods.

We evaluate the submissions and the baseline systems for the two tasks using BLEU,[32] TER,[33] chrF,[34] (all with SacreBLEU) COMET,[35] and BertScore.[36] The final ranking is done according to the average ranking of the individual metrics per family, ties on individual metrics are considered.

We report the results for Task 1 in Table 37 and for Task 2 in Table 38. M2M-100 resulted in a very strong baseline for North-Germanic languages. EdinSaar systems are second and third, followed by UBCNLPs. The ranking is consistent

[32] BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+ +version.1.4.14

[33] TER+tok.tercom-nonorm-punct-noasian-uncased+version.1.4.14

[34] chrF2+numchars.6+space.false+version.1.4.14

[35] wmt-large-da-estimator-1719 model(comet=0.1.0)

[36] bert-base-multilingual-cased_L9_no-idf_version=0.3.9(hug_trans=4.9.0.dev0)

	ca2it					ca2oc				
	BLEU	**TER**	**chrF**	**COMET**	**BertSc**	**BLEU**	**TER**	**chrF**	**COMET**	**BertSc**
M2M-100	46.6	0.390	0.694	0.743	0.913	40.2	0.405	0.673	0.341	0.892
mT5-dFT	30.4	0.551	0.571	0.235	0.872	40.1	0.395	0.680	0.402	0.897
BSC-P	42.0	0.420	0.670	0.651	0.908	57.1	0.272	0.780	0.514	0.929
CUNI-C	49.5	0.366	0.714	**0.813**	0.916	**67.1**	**0.201**	**0.832**	**0.724**	**0.952**
CUNI-P	**50.5**	**0.360**	**0.717**	0.810	**0.917**	66.9	0.202	0.829	0.719	0.951
TenTrans-C	44.1	0.410	0.680	0.667	0.912	56.1	0.309	0.813	0.617	0.941
TenTrans-P	43.2	0.418	0.671	0.640	0.910	56.5	0.304	0.817	0.640	0.944
UBCNLP-C	25.7	0.574	0.539	-0.263	0.844	51.7	0.316	0.736	0.259	0.905
UBCNLP-P	35.1	0.477	0.622	0.391	0.886	59.9	0.254	0.787	0.538	0.928

	ca2ro				
	BLEU	**TER**	**chrF**	**COMET**	**BertSc**
M2M-100	**33.1**	**0.640**	**0.535**	0.159	0.831
mT5-dFT	17.3	0.830	0.407	-0.461	0.784
BSC-P	24.9	0.695	0.490	-0.076	0.814
CUNI-C	31.8	0.644	0.533	**0.169**	**0.835**
CUNI-P	32.8	**0.640**	**0.535**	0.168	0.834
TenTrans-C	30.2	0.661	0.517	0.047	0.830
TenTrans-P	30.2	0.664	0.516	0.047	0.829
UBCNLP-C	8.6	0.884	0.311	-1.119	0.725
UBCNLP-P	11.2	0.855	0.354	-0.908	0.749

Table 40: Automatic evaluation per language pair in the Romance family of the multilingual low-resource translation task (Task 2). Best scores boldfaced. Notice that the final ranking is done per family and not per language pair as shown in Table 38.

across metrics. The quality of the second baseline, the finetuned version of mT5, is low as compared to the other systems because it has only been trained for machine translation with 5,500 parallel sentences for the 6 language pairs. EdinSaar-Primary is also a version of mT5 finetuned with 1.2 M parallel sentences and that improves translation quality significantly, but still, it lies below the multilingual baseline system trained with huge amounts of parallel data, M2M-100.

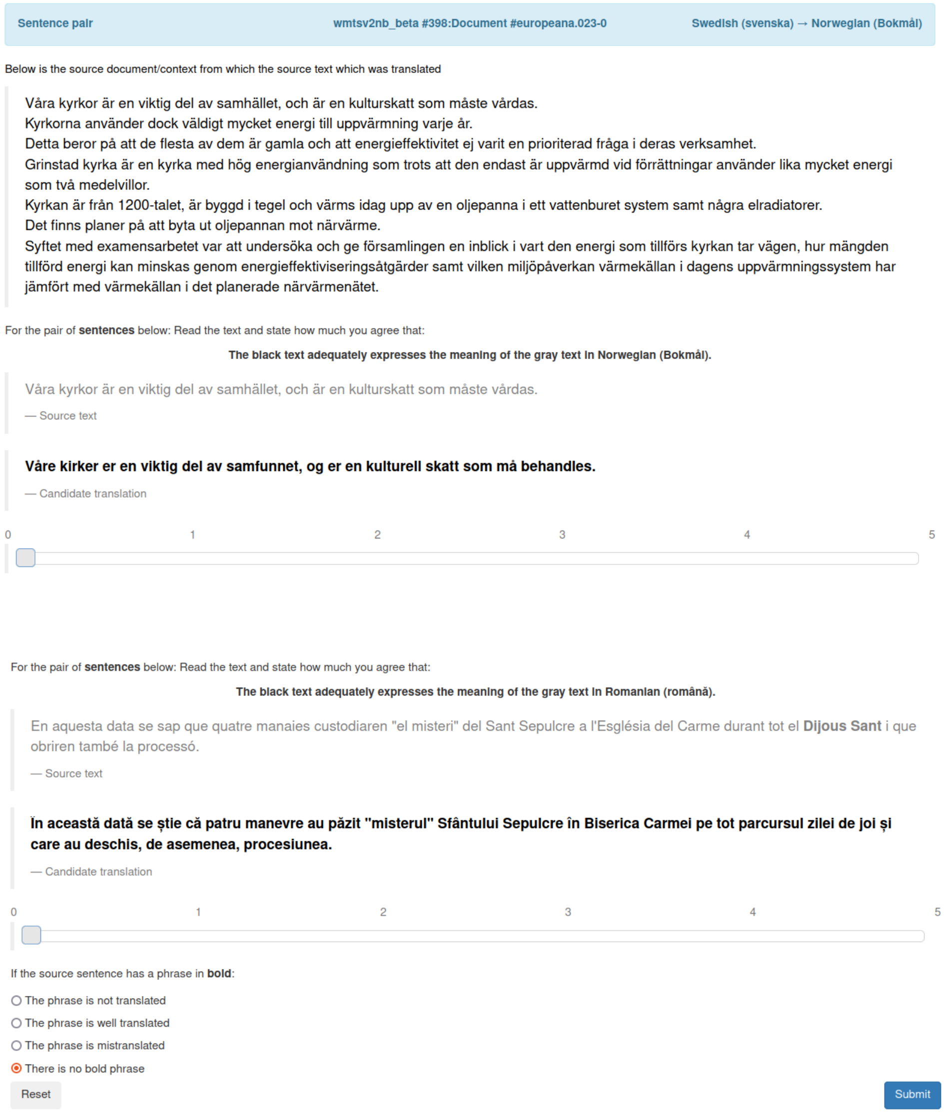
A more fine-grained analysis (Table 39) shows that translation into Icelandic is difficult for all the systems, and also translation from Icelandic into Swedish (Norwegian) is more difficult than translation from Norwegian (Swedish) into Swedish (Norwegian). Systems do not behave consistently across language pairs: mT5-devFinetuned (mT5-dFT in the table) achieves top performance when translating from Icelandic but performs poorly for the remaining pairs; UBCNLP-Contrastive (UBCNLP-C) is specially good for translating from Swedish.

For Task 2, the Romance family, the CUNI systems are significantly better than the rest, both at family and language pair levels (Tables 38 and 40). Only for ca2ro, M2M-100 is better according to some metrics; however, this system performs comparatively bad for ca2it. TenTrans and BSC perform very close one to each other. Globally, TenTrans performs better with BSC showing good performance for ca2oc. For this language pair, the reranking strategy via a character-based model by CUNI achieves very good results.

6.4 Human Evaluation

In order to complement and corroborate the automatic evaluation, we also perform human evaluation on a subset of the languages. However, since not all language pairs are covered, we cannot use the manual evaluation results for the official ranking of the systems.

The type of evaluation has been conditioned by the number and expertise of the raters we could attract. We hired a total of 14 raters: 5 Swedish annotators to rate nb2sv and is2sv documents; 3 bilingual Catalan–Occitan annotators to rate ca2oc documents and 6 bilingual Catalan–Italian annotators to rate ca2it documents. With these numbers in mind, we decided to do ratings on a Likert-like scale but following the philosophy of direct assess-



(a)

(b)

Figure 8: Modifications to the Appraise Evaluation Framework (Federmann, 2018) for the multilingual low-resource translation task. (a) We conduct reference document-level direct assessments on a discrete scale [1,5]. (b) For languages where we can conduct source document-level assessments, we we also evaluate term translation (dates and locations).

ments (DAs). We do source DA for Italian and Occitan, and reference DA for Swedish.

Following the conclusions in (Graham et al., 2020) and (Castilho et al., 2020), we perform sentence level evaluation with document context. Figure 8(a) shows that evaluators rate each sentence in context and when all the sentences in document are evaluated, the whole document is also scored. The evaluation is done using the Appraise Evaluation Framework (Federmann, 2018) with several modifications. Appraise implements document direct assessments as used in the WMT News Task evaluation campaign (Barrault et al., 2020). In our case, we have fewer annotators so we cannot expect > 15 ratings per sentence to get statistically significant results with a 100 points DA scale. To tackle this limitation, we constrain the DA scale to a 5 points Likert-like scale [1,5]. This resembles an adequacy+fluency evaluation where raters still answer the question "*The black text adequately expresses the meaning of the gray text.*", but they do not evaluate adequacy and fluency separately. After a small pilot experiment (see below), the guidelines to the evaluators were the following:

Rank a sentence with a 5 if it completely expresses the same meaning as the source/reference. Notice that we do not ask for a literal translation but for a sentence that preserves the meaning and it is grammatically correct. For a 3 score, the sentence should convey part of the meaning *of the original sentence but some relevant parts are missing or not well translated. For a 4, only non-relevant parts are not OK. For a 2, most of the sentence is wrong but still some bits, probably non-relevant, are well translated. Finally, rate the sentence with a 1 if none of the content is preserved.*

Bilingual raters allow us to do a small term translation evaluation for Catalan to Italian and Occitan. Figure 8(b) shows that we boldface some terms in the source text and evaluators are asked to say if *(i) The phrase is not translated*, *(ii) The phrase is well translated* or *(iii) The phrase is mistranslated*.

6.4.1 Data Preparation

We select test documents or parts of them to cover 100 sentences per language. Table 36 shows that considering full documents would limit the evaluation to very few texts so we select a subset of contiguous sentences in documents to make the evaluation more heterogeneous. For Catalan to Italian and Occitan, we selected fragments in 9 documents with lengths between 5 and 15 sentences; for Icelandic to Swedish fragments in 7 documents with lengths between 8 and 20 sentences; and for Norwegian to Swedish fragments in 7 documents with lengths between 7 and 22 sentences.

We extract the same 100 sentences from the participants primary submissions and from the reference. For source DA evaluation (Catalan and Occitan), the reference is also rated and used to establish human performance. For reference DA (Swedish), the reference is just used for rating translations.

Finally, we mark 60 of the source sentences in Catalan with one term each. Selected terms[37] are mostly named entities (dates, locations or titles) and might be multi-word. Named entities that appear only a few times in training data are a challenge for neural systems, so the aim is to check the quality of these translations. Since professional translators did not receive any instructions on how to translate these terms, we can observe a mixture of untranslated and translated named entities, which makes it difficult to assess its quality in an automatic way.

	nb2sv		is2sv	
System	z-score	raw	z-score	raw
M2M-100	0.7±0.6	4.2±0.8	0.1±1.0	2.0±1.1
EdinSaar	0.2±0.7	3.6±1.1	-0.1±0.8	1.9±1.0
UBCNLP	0.2±0.8	3.5±1.2	-0.4±1.0	1.6±1.1
mT5-dFT	-1.2±0.7	1.5±1.1	0.4±1.1	2.4±1.2

Table 41: Average DA and standard deviation of raw- and z-scores for all primary submissions of Task 1 in the language pairs manually evaluated.

	ca2it		ca2oc	
System	z-score	raw	z-score	raw
HUMAN	0.8±0.4	4.8±0.6	0.8±0.7	4.0±1.0
CUNI	0.5±0.7	4.4±0.9	0.5±0.8	3.6±1.1
M2M-100	0.4±0.7	4.2±1.0	-0.7±0.8	2.0±1.0
TenTrans	0.0±0.8	3.8±1.1	0.3±0.8	3.4±1.2
BSC	-0.1±0.8	3.7±1.1	0.3±0.9	3.4±1.2
UBCNLP	-0.5±1.0	3.1±1.3	0.0±0.9	3.0±1.2
mT5-dFT	-1.2±0.9	2.3±1.2	-1.0±0.7	1.7±0.9

Table 42: Average DA and standard deviation of raw- and z-scores for all primary submissions of Task 2 in the language pairs manually evaluated. HUMAN refers to the evaluation of the reference.

6.4.2 Pilot Experiment

We prepared a pilot experiment with two goals: (i) provide some training to the raters and (ii) check the feasibility of the task. For this, we prepared a manual with instructions to work with the modified Appraise interface and the guidelines for rating the translations. We populate the task with 20 translated sentences from one of the submissions. Sentences come from two test documents so that the annotators go through the full document annotation process twice.

After the pilot, we made the guidelines more concrete to accommodate the raters questions. These annotations are discarded for the final analysis described in the next section.

[37] List of terms which translation is evaluated manually: Plaça del Mercadal, segle XV, segle XIX i XX, la Casa Pinyol, Festes de Maig, Rambla de Badalona, la Cremada, la Segona República, Josep Maria Cuyàs, Baró de Maldà, 11 de maig de 1940, Francesc de Paula Giró i Prat, Aristeus antennatus, Productes de l'Empordà, 400 metres, mitjan segle XX, Canyó de Palamós, Confraria de Pescadors de Palamós, finals del segle XIX, Xat de Benaiges, començaments del segle XX, "salvitxada", la calçotada, Alt Camp, Congrés de Cultura Catalana, Valls, Concurs de salsa de la "calçotada", Fogueres de Sant Antoni, Nadal, Sant Antoni, Química Orgànica, Universitat de Barcelona, Junta d'Energia Nuclear, Universitat de Chicago, Universitat de València, Física Teòrica, Mecànica Teòrica, Premi d'Investigació Ramón y Cajal, Manaies de Girona, any 1751, Dijous Sant, Setmana Santa, segles xviii i xix, 1851, mitjans de segle XIX, finals del XVIII, port del Masnou, dos quilòmetres i mig, Club Nàutic del Masnou, Creu Roja, festival Ple de Riure, Masnou, N-II, Premià de Mar, any 2019, platja d'Ocata, Michelin, Ferran Adrià, El Celler de Can Roca, Can Fabes

6.4.3 Results

The results of the evaluation task are the average DA scores per system. In order to take into account that some raters might be more strict than others, we rank the systems according to the z-score, where the DA score is mean-centered and normalised per rater.

Inter-annotator agreement as measured by Fleiss' κ (Fleiss, 1971) is moderate: 0.32±0.03 (nb2sv, fair agreement), 0.16±0.04 (is2sv, slight agreement), 0.28±0.03 (cat2it, fair agreement) and 0.16±0.02 (ca2oc, slight agreement). These values are in agreement with previous analyses (Castilho, 2020). Intra-annotator agreement ranges from 0.88±0.06 to 0.24±0.09 for the North-Germanic languages and from 0.56±0.09 to -0.04±0.07 for the Romance family. We discard raters with $\kappa \sim 0$ and report results with 4 raters for Swedish, 3 for Catalan–Occitan and 4 for Catalan–Italian. Tables 41 and Table 42 show the results for Task 1 and Task 2 respectively.

For Task 1, we obtain very different scores depending on the language pair. This is in line with the automatic evaluation: translations from Icelandic do not behave in the same way as Swedish and Norwegian which are closer languages. Baselines perform very well on this family, but not simultaneously. M2M-100 offers good translation quality for nb2sv while mT5-dFT is specially good for is2sv. For is2sv, systems are not statistically significantly different, for nb2sv mt5-dTF is significantly worse than the others and EdinSaar and UBCNLP show similar performance.

For Task 2, the reference (HUMAN) is ranked first in both language pairs, but the deviation is large and it is not significantly better than the CUNI system. For ca2it, HUMAN is not significantly better than the baseline system M2M-100 either. In some cases though, the distinction seemed to be easy. Raters pointed out several reasons: (i) mistranslations of very frequent words —*got* in Catalan (cup, glass) translated into Italian as *getto* (jet), *grigio* (gray) or *vetro* (glass, the material); (ii) bad translation in context of ambiguous words —*quarentena* in Catalan translates into Italian as *quarantina* (about fourty) or *quarantena* (quarantine); (ii) mistaken roots (this can be related to BPE subunits as explained below) —*calçots* (a local vegetable) translated as *calzatura* (footwear); or changing words —*un físic català* (a Catalan physicist) translated as *un fisico spagnolo* (a Spanish physicist).

System	ca2it well	ca2it mis	ca2it no	ca2it Σ	ca2oc well	ca2oc mis	ca2oc no	ca2oc Σ
HUMAN	53	0	3	56	40	0	2	42
CUNI	39	3	5	47	30	7	1	38
M2M-100	33	2	6	41	26	9	0	35
TenTrans	37	0	9	46	32	4	1	37
BSC	27	7	5	39	33	4	0	37
UBCNLP	29	16	1	46	19	1	0	20
mT5-dFT	20	17	10	47	25	11	4	40

Table 43: Number of **well** translated, **mis**-translated and **not** translated terms for the language pairs manually evaluated for Task 2. The last column per language shows the total number of terms considered from the maximum of 60 bold faced terms (see text).

Similar to the automatic evaluation, TenTrans and BSC are very close to each other according to the human ratings although the two architectures are completely different. The evaluation also confirms the bad performance of M2M-100 on ca2oc but its good performance on ca2it. In general, all the systems perform worse on ca2oc than ca2it according to the raw scores in Table 42, but the trend is reversed when analysing the z-scores. This result points to differences between the scale that annotators used in the two tasks even if they received the same instructions. Notice that almost all automatic metrics but COMET tend to score higher ca2oc than ca2it for most systems.

Term translation. The evaluation against the source for the Romance languages allows us to study the translation quality of selected terms. For ca2it we use the annotations from 5 raters but only 2 were considered for ca2oc as the remaining raters did not do the task properly. The agreement for this task is 0.34±0.05 (ca2it) and 0.19±0.05 (ca2oc). Table 43 shows the number of well translated, mis-translated and untranslated terms for both pairs.

For each term, we sum the votes from all the raters per class (well translated, mis-translated or untranslated) and consider the winning class the one with the majority of votes. In case there is a tie with 2 or more classes, the term is not considered in the analysis, this is why the last columns Σ in Table 43 differ from 60. The disagreement is high, and one of the causes is the ambiguity in the annotation of toponyms. For instance, the name of the city of "Valls" has been evaluated 17 times: 7 times as well translated and 10 times as not translated being always the translation

"Valls". The same happens with other toponyms and years. This ambiguity damages specially the majority voting for Occitan (low Σ) since we only consider 2 raters.

The systems with the largest number of mistranslations are those with less access to the task languages, that is, the baselines. mT5-devFinetuned and M2M-100 (specially for Occitan) do the most mistakes. A curious case is UBNLP which only produces 1 mistranslation for Occitan but 16 for Italian. Also BSC generates more errors for Italian (7) than for Occitan (4) even though translation quality into Italian is higher than into Occitan. Looking at some examples, we hypothesise that this can be related to the subunit segmentation strategy. For instance, the word "calçotada" is translated as *calzotada, calzolata* or as we have seen before *calzatura* in Italian, where no Italian word for this concept exists. For Occitan, it is always translated by *calçotada* (BPE units in Catalan and Occitan might be the same, but not for Italian), only two times it is mistranslated as *escòla*.

Besides these errors that might be due to the split in subunits, we also observe multi-word named entities where one of the words has been literally translated and the others have not. Also, in few occasions, a number (specially centuries) is translated by another one.

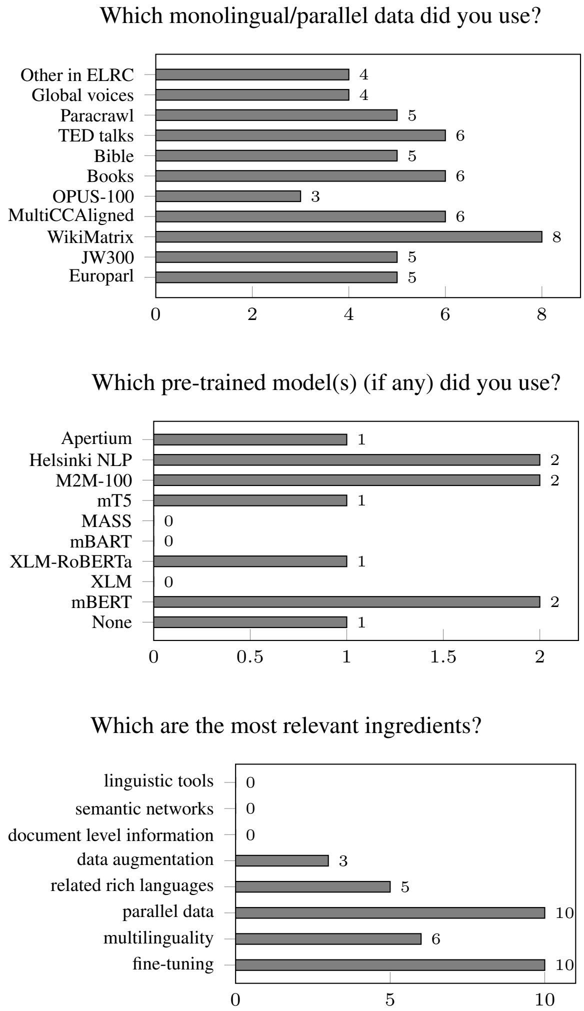


Figure 9: Resources used by the participants to train the systems submitted to the multilingual low-resource translation task (10 responses).

6.5 Discussion

This shared task faced three challenges: multilingual translation, document translation and in-domain (cultural heritage) translation. 60% of the submissions approached multilinguality with a single system while 40% used a combination of several bilingual systems. None of the participants focused on the document-level aspect of the task, and those who dealt with the specific domain did not use any of the in-domain multilingual lexicons but selected in-domain data from the available training corpus.

More details and comparisons among the submissions can be found in Figures 9 and 10. Figure 9 focuses on the resources. Participants did not use all the data available, probably because of its heterogeneous nature and the difference of language pairs available in the different corpora. WikiMatrix is the favourite corpus, with 80% of the submissions trained on it. 90% of the systems used some kind of pretrained model: from language models such as mBERT (TenTrans, EdinSaar) or XLM-RoBERTa (BSC) to machine translation models such as M2M-100 (TenTrans) or Helsinki's NLP (UBCNLP). There is no clear favourite system here, and each team followed a different approach. In all cases, systems were finetuned with language specific data, either data made available for the task or backtranslations made by themselves. 50% of the submissions also used data from the related high resourced languages for training.

Figure 10 compares the architectures. As expected, neural systems dominate the number of submissions. In fact, all of them where 100% neural, without any hybridisation with any non-neural component. All participants used direct translation, either multilingual (60%) or bilingual (40%), but none of them submitted translations done

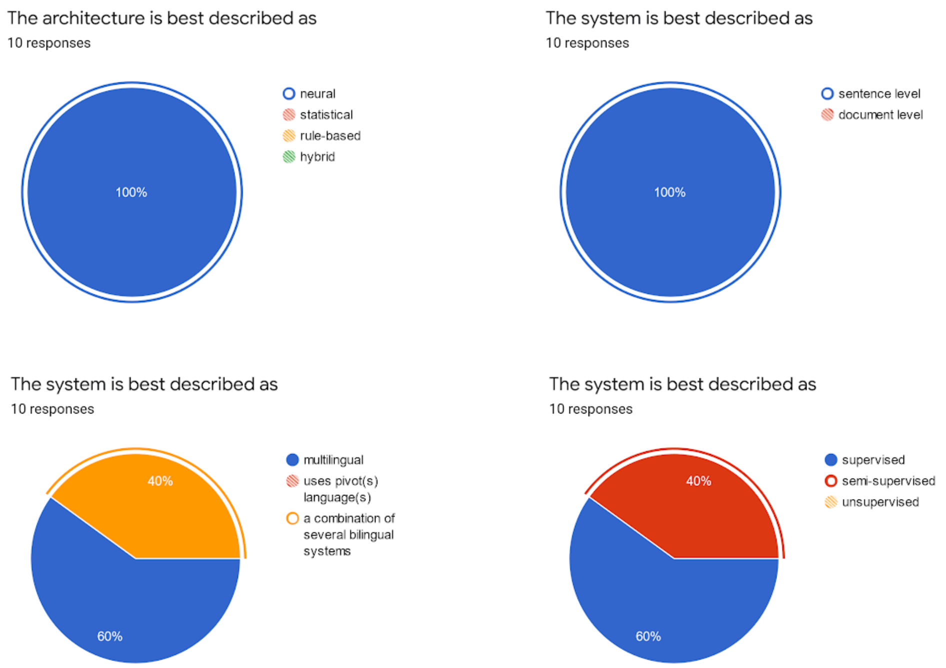


Figure 10: Main characteristics of the systems submitted to the multilingual low-resource translation task. Percentages are over the sample of 10 submissions.

through a pivot language. One team, CUNI, tried pivot through English for the Romance languages but translation quality was significantly better with direct systems. TenTrans used a pivot language for creating a synthetic corpus using backtranslation. Similarly to CUNI's, the approach worked well for ca2it and ca2ro but did not work at all for the lowest resourced language, Occitan, damaging the quality of the multilingual system as a whole. In both cases, multilingual systems trained with parallel data of the task languages plus additional corpora with the related rich languages as source gave the best performance.

Data augmentation via backtranslations and/or parallel data including high-resourced languages have been beneficial for all the systems. Two teams also got improvements by selecting data close to the domain of the validation set, but the in-domain adaptation was not decisive to win the shared task. TenTrans extracted in-domain sentences with a domain classifier trained on mBERT in Task 2 while EdinSaar used cross-entropy for the same purpose in Task 1.

In this shared task, we have evaluated systems per family, but differences among translation pairs are significant and determine the final ranking. The trends for the 2 families are similar. One of the languages has a relatively large amount of data (Swedish/Italian), the second language in terms of amount of data is the most distant one within the family (Icelandic/Romanian) and the lowest-resourced language is linguistically very similar to the richest language (Norwegian Bokmål/Occitan). Icelandic is the bottleneck for Task 1 and Romanian for Task 2 showing that in this case the distance between languages is more important than the amount of data.

It is interesting to see how the ranking depends on the language pair. The most extreme case is our baseline mT5-devFinetuned which performed the best when translating from Icelandic and the worst in the other cases (Task 1). Similarly but not so extreme, UBCNLP-Contrastive performed very well when translating from Swedish and significantly worse on the other cases. In Task 2, Romance languages, the two baselines specially M2M-100, are penalised by the bad performance on ca2oc showing that the amount of Occitan text might be too diluted in their multilingual training. M2M-100 is the best for ca2ro, and this is the only pair where the best system is not CUNI. For all the systems, ca2ro is the most difficult pair.

Finally, we want to emphasise the correlation between automatic and human evaluations among systems even though standard deviations are high and top performing systems are not significantly different.

7 Automatic Post Editing

This section presents the results of the 7^{th} round of the WMT task on MT Automatic Post-Editing. The task consists in automatically correcting the output of a "black-box" machine translation system by learning from human-revised machine-translated output. In continuity with last year, the challenge consisted of fixing the errors present in English Wikipedia pages translated – into German and Chinese – by state-of-the-art, not domain-adapted neural MT (NMT) systems unknown to participants. Despite a number of data downloads in line with the previous rounds, this year we observed an unexpected drop in participation: two teams participated in the English-German task, submitting two runs each, while the English-Chinese task had no participants. Most likely, this setback can be ascribed to the difficulty to handle the released test data, which are characterized by NMT output of very high quality. This is reflected by much higher baseline results compared to last year (18.05 TER / 71.07 BLEU for en-de, 22.73 TER / 69.2 BLEU for en-zh), which only one run was able to improve according to both the automatic metrics used (-0.77 for the primary TER metric and +0.48 for the secondary BLEU metric). Nevertheless, the outcomes of human evaluation still reveal the ability of APE systems to improve MT output quality: significant gains over the baseline are indeed observed for all the participating systems.

7.1 The Task

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:

- Improve MT output by exploiting information unavailable to the decoder, or by performing deeper text analysis that is too expensive at the decoding stage;

- Cope with systematic errors of an MT system whose decoding process is not accessible;

- Provide professional translators with improved MT output quality to reduce (human) post-editing effort;

- Adapt the output of a general-purpose MT system to the lexicon/style requested in a specific application domain.

This 7^{th} round of the WMT APE shared task kept the same overall evaluation setting of the previous six rounds. Specifically, the participating systems had to automatically correct the output of an unknown "black box" (neural) MT system by learning from training data containing human revisions of translations produced by the same system. The selected language pairs (English-German and English-Chinese) and the data domain (Wikipedia articles) were the same of last year (Chatterjee et al., 2020), as well as the type of MT systems (generic NMT systems not adapted to the target domain).

7.2 Data, Metrics, Baseline

7.2.1 Data

In continuity with all previous rounds, participants were provided with **training** and **development** data consisting of (*source*, *target*, *human post-edit*) triplets (7,000 for the training and 1,000 for the development sets for both languages) where:

- The source (SRC) is a tokenized English sentence;

- The target (TGT) is a tokenized German/Chinese translation of the source, which was produced by a generic, black-box NMT system unknown to participants.[38]

- The human post-edit (PE) is a tokenized manually-revised version of the target, which was produced by professional translators.

For the English-German sub-task, two additional training resources were made available to participants. These are: *i)* the corpus of 4.5 million artificially-generated post-editing triplets described in (Junczys-Dowmunt and Grundkiewicz, 2016), and *ii)* the 14.5 million artificially-generated instances of the English-German section of the eSCAPE corpus (Negri et al., 2018).

[38] The NMT systems for both the languages are based on the standard Transformer architecture (Vaswani et al., 2017) and follow the implementation details described in (Ott et al., 2018). They were trained on publicly available MT datasets including Paracrawl (Bañón et al., 2020) and Europarl (Koehn, 2005), summing up to 23.7M parallel sentences for English-German and 22.6M for English-Chinese.

Test data consisted of newly-released (*source*, *target*) pairs (1,000 in total for each target language), similar in nature to the corresponding elements in the train/dev sets (i.e. same domain, same NMT architectures). The human post-edits of the target elements were left apart to measure APE systems' performance both with automatic metrics (TER, BLEU) and via manual assessments.

7.2.2 Metrics

Also this year, the participating systems were evaluated both by means of automatic metrics and manually (see Section 7.5). Automatic evaluation was carried out by computing the distance between the automatic post-edits produced by each system for the target elements of the test set, and the human corrections of the same test items. Case-sensitive TER (Snover et al., 2006) and BLEU (Papineni et al., 2002) were respectively used as primary and secondary evaluation metrics. The official systems' ranking is hence based on the average TER calculated on the test set by using the TERcom[39] software: lower average TER scores correspond to higher ranks. BLEU was computed using the multi-bleu.perl package[40] available in MOSES. Automatic evaluation results are presented in Section 7.5.1.

Manual evaluation was conducted via source-based direct human assessment (Graham et al., 2013). Complete details are provided in Section 7.5.3.

7.2.3 Baseline

Also this year, the official baseline results were the TER and BLEU 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. For each submitted run, the statistical significance of performance differences with respect to the baseline was calculated with the bootstrap test (Koehn, 2004).

7.3 Complexity indicators

To get an idea of the difficulty of the task, in previous rounds we have focused on three aspects of the released data, which provide us with information about the possibility of learning useful correction patterns during training and successfully applying them at test time. These are: *i)* repetition rate, *ii)* MT quality, and *iii)* TER distribution in the test set. For the sake of comparison across the seven rounds of the APE task (2015–2021), Table 44 reports, for each dataset, information about the first two aspects. The third one, instead, will be discussed by referring to Figure 11. Concerning this year's round, we only report information for the English-German sub-task, the only one for which we had participants; also the discussion henceforth will exclusively focus on this sub-task.

7.3.1 Repetition Rate

The repetition rate, 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. Larger values indicate a higher text repetitiveness that may suggest a higher chance of learning from the training set correction patterns that are applicable also to the test set. However, over the years, the influence of repetition rate in the data on systems' performance was found to be marginal.[41] For the sake of completeness, we hence just observe that, being drawn from the same Wikipedia domain, this year's data feature very low repetitiveness values (i.e. 0.73, 0.78, and 0.76 respectively for the SRC, TGT and PE elements), which are comparable to those from last year (0.653, 0.823, and 0.656). In spite of this, while last year's gains over the baseline were the highest ever observed in the APE task history, this year's results are significantly lower. This suggests the higher importance of other complexity factors, on which repetition rate might have an additive effect that still has to be fully understood.

7.3.2 MT Quality

MT quality, that is the initial quality of the machine-translated (TGT) texts to be corrected, is indeed a much more reliable indicator of task difficulty. We measure it by computing, the TER (↓) and BLEU (↑) scores using the human post-edits as reference. As discussed in (Bojar et al., 2017; Chatterjee et al., 2018, 2019, 2020) higher quality of the original translations leaves to the APE systems a smaller room for improvement since they have, at the same time, less to learn during training and less to correct at test stage. On one

[39] http://www.cs.umd.edu/~snover/tercom/

[40] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl

[41] The analyses carried out over the years produced mixed outcomes, with impressive final results obtained in spite of low repetition rates (Chatterjee et al., 2020) and vice-versa (Chatterjee et al., 2018, 2019).

	Lang.	Domain	MT type	RR_SRC	RR_TGT	RR_PE	Baseline BLEU	Baseline 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

Table 44: 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 row (δ TER) indicates, for each evaluation round, the difference in TER between the baseline (i.e. the "*do-nothing*" system) and the top-ranked submission. For this year's round we report results for the only sub-task – English-German – for which we had participants.

side, training on good (or near-perfect) automatic translations can drastically reduce the number of learned correction patterns. On the other side, testing on similarly good translations can *i)* drastically reduce the number of corrections required and the applicability of the learned patterns, and *ii)* increase the chance to introduce errors, especially when post-editing near-perfect TGTs. The findings of all previous rounds of the task support this observation and, as discussed in Section 7.5, this year is no exception. For English-German, the quality of the initial translations (18.05 TER / 71.07 BLEU) is close the level of the "hardest" previous rounds (2017-2019), characterized by baseline scores in the 15.5-16.8 TER interval (and BLEU>70.0). Accordingly, this year's gains over the baseline amount to less than 1 TER/BLEU points. The strict correlation between the quality of the initial translations and the actual potential of APE is hence confirmed.

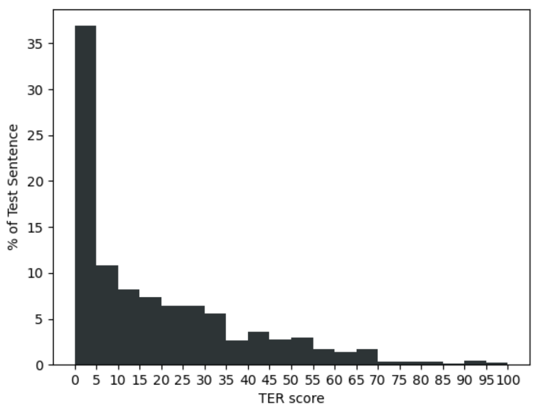


Figure 11: TER distribution in the **English-German** test set.

7.3.3 TER Distribution

A third reliable complexity indicator is the TER distribution (computed against human references) for the translations present in the test sets. Although TER distribution and MT quality can be seen as two sides of the same coin, it's worth remarking that, even at the same level of overall quality, more/less peaked distributions can result in very different testing conditions. Indeed, as shown by previous analyses, harder rounds of the tasks were typically characterized by TER distributions particularly skewed towards low values (i.e. a larger percentage of test items having a TER between 0 and 10). On one side, the higher the proportion of (near-)perfect test instances requiring few edits or no corrections at all, the higher the probability that APE systems will perform unnecessary corrections penalized by automatic evaluation metrics. On the other side, less skewed distributions can be expected to be easier to handle as they give to automatic systems a larger room for improvement (i.e. more test items requiring - at least minimal - revision). In the lack of more focused analyses on this aspect, we can hypothesize that, in ideal conditions from the APE standpoint, the peak of the distribution would be observed for "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.[42]

Also with respect to this complexity indicator, this year's test set looks particularly difficult to handle. As shown in Figure 11, more than 35%

[42]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.

ID	Participating team
PVIE	Amazon Prime Video, India (Sharma et al., 2021)
Netmarble	Netmarble AI Center, South Korea Korea (Oh et al., 2021)

Table 45: Participants in the WMT21 Automatic Post-Editing task.

of the test instances feature a TER between 0 and 5 and almost 50% of them have 0<TER<10. This distribution, which is very different from last year (where less than 7% of the test samples had 0<TER<5 and ∼55% of them had 15<TER<45), is similar to the one featured by the most challenging datasets from previous rounds.

All in all, the small gains over the baseline mentioned above also confirm the strict correlation between TER distribution and task difficulty. This goes hand in hand with the above considerations about MT quality and, together with the possible additive effect of very low repetition rate values in raising the difficulty bar, might have discouraged potential participants.

7.4 Submissions

As shown in Table 45, we received submissions from two teams, which is indeed a significant drop with respect to last year's round. Moreover, as anticipated, both teams participated only in the English-German sub-task by submitting 2 runs each.

Amazon Prime Video (*PVIE*). Amazon participated with a model leveraging a state-of-the-art MT system based on fairseq (Ott et al., 2019) and pre-trained on data from the WMT'19 News Translation task (Barrault et al., 2019). The basic model is first fine-tuned on the APE dataset, by creating (*source*, *target*) pairs where the *source* is a concatenation of the SRC and MT elements of the APE data and the *target* is the corresponding PE element. Then, to cope with the domain mismatch between the initial training data and the APE task ones, the model is fine-tuned on *i)* data drawn from WikiMatrix (Schwenk et al., 2019) (64k parallel sentences after cleaning), *ii)* additional APE samples (45k triplets) from previous rounds (2016-2018) of the shared task, and *iii)* this year's APE data. The primary submission is obtained by ensembling models built from different combinations of the available data.

Netmarble AI Center (*Netmarble*). Netmarble participated with a Transformer-based system (Vaswani et al., 2017) built using: *i)* the WMT21 News Translation data, *ii)* the additional artificial synthetic data provided to the APE task participants, and *iii)* data augmentation techniques that make use of an external MT component. These resources are processed through a curriculum training procedure aimed to step-wise learn from easier problems to more complex ones. Multi-task learning is also applied to alleviate data sparsity issues by sharing knowledge across related tasks (in this case part of speech recognition, named entity recognition, masked language modeling and keep/translate classification). All tasks are jointly trained and, to cope with imbalanced data from the selected tasks, task-specific losses – namely focal loss (Lin et al., 2017) and class-balanced loss (Cui et al., 2019) - are exploited in addition to standard cross-entropy. Moreover, dynamic weight average (Liu et al., 2019), which adapts the task weighting over time by considering the rate of change of the loss for each task, is applied to optimize the contribution of each task in the multi-task framework.

7.5 Results

7.5.1 Automatic evaluation

Participants' results are shown in Table 46. The submitted runs are ranked based on the average TER (case-sensitive) computed using human post-edits of the MT segments as reference, which is the APE task primary evaluation metric. We also report the BLEU score, computed using the same references, which represents our secondary evaluation metric.

As it can be seen from the table, the two rankings slightly differ: while the top submission (17.28 TER, 71.55 BLEU) is the same, the BLEU-based ranking presents few swaps, with the *do nothing* baseline reaching the 2nd position. One obvious observation is that these fluctuations are due to the fact that all systems substantially perform on par: except for one case (i.e. the 0.77 TER reduction achieved by the top submission), all the results' differences with respect to the baseline are indeed not statistically significant.

Quite surprisingly, we also observe that the best

		TER	BLEU
en-de	Netmarble_CURRICULUM-ENSEMBLE_CONTRASTIVE	**17.28**	71.55
	PVIE_single_CONTRASTIVE	17.74	70.54
	PVIE_ensemble_PRIMARY	17.85	70.5
	Netmarble_CURRICULUM-MTL_PRIMARY	17.97	70.53
	Baseline	18.05	71.07

Table 46: Results for the WMT21 APE English-German – average TER (↓), BLEU score (↑) Statistically significant improvements over the baseline are marked in **bold**.

submission for both participants is the contrastive one. This highlights the difficulty to select the best configuration during system development, and indirectly confirms the difficulty to handle APE data characterized by very high MT quality, TER distribution skewed towards perfect/near-perfect translations and very low repetition rate values.

7.5.2 Systems' behaviour

Modified, improved and deteriorated sentences. In light of the hard conditions posed by what seems to be the hardest APE dataset ever released, we now turn an eye toward the changes made by each system to the test instances. To this aim, Table 47 shows, for each submitted run, the number of modified, improved and deteriorated sentences, as well as the overall system's precision (i.e. the proportion of improved sentences out of the total number of modified instances for which improvement/deterioration is observed). It's worth noting that, as in the previous rounds, the number of sentences modified by each system is higher than the sum of the improved and the deteriorated ones. This difference is represented by modified sentences for which the corrections do not yield any TER variations. This grey area, for which quality improvement/degradation can not be automatically assessed, contributes to motivate the human evaluation discussed in Section 7.5.3.

As it can be seen from the table, systems' behaviour reflects the difficulty to handle this year's test set. The quite low percentage of modified sentences (50.2 on average, 46.2 for the top submission) is in line with our previous observations about TER distribution (see Section 7.3.1). With ∼50% of the test instances having 0<TER<10, all systems seem to have properly managed the small room for intervention by not exceeding the number of expected corrections. Accordingly, different from last year,[43] systems' final scores are inversely proportional to their aggressiveness.

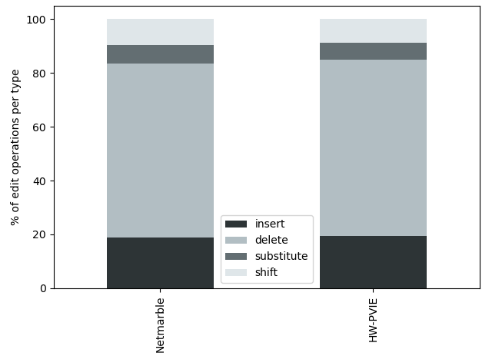


Figure 12: Distribution of edit operations (insertions, deletions, substitutions and shifts) performed by the two primary submissions to the English-German task.

Precision-wise, however, we are far from last year's values: despite lower aggressiveness, system's precision is 51.12 on average (in 2020 it was 58.0) with the best run peaking at 53.96 (vs 69.0 in 2020). This is due to significant variations in the percentage of improved (43.5 on average, 45.67 for the top submission) and deteriorated sentences (41.6 on average, 38.96 for the winning system), which are very different from last year where, on a simpler test set, the average values were respectively 58.2 and 23.6.

Edit operations. Similar to previous rounds, we analysed systems' behaviour also in terms of the distribution of edit operations (insertions, deletions, substitutions and shifts) done by each system. This fine-grained analysis of how systems corrected the test set instances is obtained by computing the TER between the original MT output and the output of each primary submission taken as reference. Similar to last year, and in line with the close TER/BLEU results obtained by the two systems, differences in their behaviour are barely visible. Both of them are characterised

[43] On the much simpler 2020 test set, featuring only ∼15.0% of instances with 0≤TER≤10, the modified sentences were 69.2% on average, with the more aggressive behaviour of the top systems peaking to more than 90.5%.

Systems	Modified	Improved	Deteriorated	Prec.
Netmarble_CURRICULUM-ENSEMBLE_CONTRASTIVE	462 (46.2%)	211 (45.67%)	180 (38.96%)	53.96
PVIE_single_CONTRASTIVE	504 (50.4%)	212 (42.06%)	212 (42.06%)	50.0
PVIE_ensemble_PRIMARY	508 (50.8%)	215 (42.32%)	218 (42.91%)	49.65
Netmarble_CURRICULUM-MTL_PRIMARY	533 (53.3%)	235 (44.09%)	227 (42.59%)	50.87
Average	50.2	43.5	41.6	51.12

Table 47: Number (raw and proportion) of test sentences modified, improved and deteriorated by each run submitted to the APE 2021 **English-German** sub-task. The "Prec." column shows systems' precision as the ratio between the number of improved sentences and the number of modified instances for which improvement/deterioration is observed (i.e. Improved + Deteriorated).

	Avg	Avg z
Netmarble_CURRICULUM-MTL_PRIMARY	79.82	0.144
Netmarble_CURRICULUM-ENSEMBLE_CONTRASTIVE	78.52	0.095
PVIE_ensemble_PRIMARY	76.85	0.02
PVIE_single_CONTRASTIVE	76.67	0.011
test.mt	69.68	-0.27

Table 48: Results for the WMT21 APE **English-German – human evaluation**. Systems ordered by DA score; systems within a cluster are considered tied; lines indicate clusters according to Wilcoxon rank-sum test $p < 0.05$.

by a large number of deletions (65.0% on average), followed by insertions (19.2%), shifts (9.2%) and substitutions (6.5%). Although this year's test set turned out to be very different in terms of difficulty, this distribution is practically identical to last year. More thorough future investigations would be needed to find clear explanations for these observations. For the time being, to get further insights about systems' performance, we now complement our analysis by discussing the outcomes of human evaluation of the submitted runs.

7.5.3 Human evaluation

In order to complement the automatic evaluation of APE submissions, manual evaluation of the 4 submissions for English-German was conducted. In this section, we present the evaluation procedure, as well as the results obtained.

7.6 Evaluation procedure

We evaluated the overall quality of the MT and PE output using source-based direct assessment (Graham et al., 2013; Cettolo et al., 2017; Bojar et al., 2018b). We used the same instructions that are used in the News Translation track of WMT2021. Instead of using crowd-workers, we hired 2 professional translators for English-German that are native German speakers as suggested by Freitag et al. (2021a).

Human evaluation results for English-German are summarized in Table 48. Similar to last year's task (Chatterjee et al., 2020), all 4 submissions significantly improved the original MT output. Furthermore, the APE system of *Netmarble_CURRICULUM-MTL_PRIMARY* significantly outperforms all other submission and can be declared as the single winner of this years' APE task. Interestingly, the human evaluation results show no correlation with the automatic scores from Table 46 which confirms the findings from Freitag et al. (2019) that automatic evaluation is hard for post-edited systems.

7.7 Summary

We presented the results from the 7^{th} shared task on Automatic Post-Editing at WMT. This round of the challenge featured the same overall setting of last year. Specifically, the language directions were the same (English-German and English-Chinese), as well as the domain of the data (Wikipedia articles) and the neural MT systems used to produce the translations to be automatically post-edited. Also the evaluation process was carried out in continuity with the past, both with automatic metrics (TER and BLEU, respectively the primary and secondary metrics) and by means of human evaluation (via source-based direct assessment, similar to the News Translation track but involving professional translators). According to several complexity indicators (repetition rate, original MT quality and TER distribution), this year's data can be safely considered as the most difficult one ever released. On one side, this might have discouraged potential participants, which were only two for the English-German sub-task. On the other side, it contributes to explain the lower results compared to last year.

Indeed, only one submitted run was able to achieve statistically significant improvements over the *do-nothing* baseline in terms of our primary automatic metric. Nevertheless, all submissions were consistently ranked higher by human evaluators, indicating the effectiveness of APE technology even under such extremely challenging conditions.

8 Conclusion

The news translation task in 2021 covered 20 translation pairs, 14 of which had English on the source or target side and 6 were without English. Direct assessment (DA) was the main golden truth again, although the style varied across language pairs. Into-English translation was evaluated against human reference translation, preserving the order of sentences in a document but not presenting the whole document at once (SR+DC). Out-of-English and some of non-English pairs offered the full document context to the annotators and allowed them to revisit the scores assigned to individual segments (SR+FD), evaluating against the source. Four non-English pairs used a simpler evaluation without any document context (SR−DC). For English→Czech, English→German and Chinese→English, a contrastive DA scoring was also tested, presenting individual sentences in pairs of candidate translations (contr:SR-DC), aimed at a more discerning pairwise comparisons. And finally, an alternative scoring style called GENIE was additionally applied to German→English.

Document context was found to be extremely important for evaluation of high-quality MT systems. The ranking of participating systems differs considerably between SR+FD and contr:SR-DC. In particular, human reference is scored well if full document context is available throughout the annotation but tends to be surpassed by top systems when sentences are evaluated in isolation. Surprising effects were also observed when using these evaluation methods on different human translations.

The triangular machine translation task encouraged participants to use all the parallel data provided (involving direct and indirect sources) to build a better machine translation system for the particular language pair and direction (Russian-to-Chinese). The participants explored several modeling choices and data augmentation strategies that would help practitioners when building machine translation systems involving non-English language pairs.

The multilingual low-resource translation task dealt with two Indo-European language families: North Germanic and Romance. The best performing systems used multilingual supervised machine translation models enriched with backtranslated data and additional sentences from higher-resourced languages in the same family. Pivot translation via these high-resourced counter-parts and in-domain data selection was not beneficial for the final performance.

The results of the task on automatic post-editing were highly influenced by the difficulty of this year's data, which can also explain a drop in participation (two teams, only in the English-German sub-task). In light of the very high quality of the translation to be automatically corrected, the very skewed TER distribution towards near-perfect translations and the very low repetition rate in the data, it comes as no surprise that only one run was able to outperform the strong *do-nothing* baseline with statistically significant improvements. Nevertheless, human evaluation results reveal significant gains by all runs, attesting the difficulty to apply automatic evaluation procedures to APE and, on a positive note, the effectiveness of the proposed methods.

Acknowledgments

The organizers of the automatic post-editing task would like to thank Apple and Google Research for their support and sponsorship in organizing this round of the APE challenge. The organizers of the triangular machine translation task would like to thank DiDi Chuxing for providing data and research time to support this shared task.

The multilingual low-resource translation for Indo-European languages task has been funded by the European Language Resource Coordination ELRC (SMART 2019/1083) and LT-BRIDGE (H2020, 952194), and supported by the Directorate-General for Language Policy, Ministry of Culture. Government of Catalonia. We are thankful to Europeana for providing source texts in Icelandic, Norwegian and Swedish and to Antonio Toral for fruitful discussions on human evaluation.

For the news task, we are very grateful to the sponsors of our test sets. Translation of the tests sets received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement Nos. 825460 and 825303 (Elitr and Bergamot, for Czech↔English) and No. 825299 (GoURMET, for Hausa↔English). The translation of the German↔English and Chinese↔English test sets was funded by Microsoft, the Russian↔English test sets were funded by Yandex, the Japanese↔English test sets by University of Tokyo and NTT and the French↔German test sets by LinguaCustodia. The Icelandic↔English task was sponsored by the Language Technology Programme for Icelandic 2019–2023. The programme, which is managed and coordinated by Almannarómur, is funded by the Icelandic Ministry of Education, Science and Culture. The Bengali↔Hindi and Xhosa↔Zulu test sets were sponsored by Facebook. The human evaluation was co-funded by Microsoft, Toloka AI, and Facebook. The effort that goes into the manual evaluation campaign each year is impressive, and we are grateful to all participating individuals and teams for their work. We are also grateful to the many workers who contributed to the human evaluation via Mechanical Turk.

The organizers of the Similar Languages Task would like to thank Pangeanic for the Spanish, Catalan, Portuguese, and Romanian data and the Directorate-General for Language Policy at the Ministry of Culture, Government of Catalonia for the Catalan translations. They further thank the AI Journal - Funding Opportunities for Promoting AI Research for supporting the French - Maninka data collection. The French - Bambara dataset is partially funded by a grant awarded by the Lacuna Fund within the scope of the program "Datasets for Languages in Sub-Saharan Africa". We also thank Andrij Rovenchak for the support on data collection. Marta R. Costa-jussà would like to acknowledge the support of the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 947657).

Ondřej Bojar would like to acknowledge the grant no. 19-26934X (NEUREM3) of the Czech Science Foundation for his time as well as co-funding manual annotation.

Support was provided by Science Foundation Ireland in the ADAPT Centre for Digital Content Technology (`www.adaptcentre.ie`) at Trinity College Dublin funded under the SFI Research Centres Programme (Grant 13/RC/2106) co-funded under the European Regional Development Fund.

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*, Online. 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*, 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*, Florence, Italy. Association for Computational Linguistics.

George Awad, Asad Butt, Keith Curtis, Jonathan Fiscus, Afzal Godil, Yooyoung Lee, Andrew Delgado, Jesse Zhang, Eliot Godard, Baptiste Chocot, Lukas Diduch, Jeffrey Liu, Alan Smeaton, Yvette Graham, Gareth Jones, Wessel Kraaij, and Georges Quenot. 2021. Trecvid 2020: A comprehensive campaign for evaluating video retrieval tasks across multiple application domains.

George Awad, Asad Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, Andrew Delgado, Jesse Zhang, Eliot Godard, Luca Diduch, Alan F. Smeaton, Yvette Graham, and Wessel Kraaij. 2019. Trecvid 2019: An evaluation campaign to benchmark video activity detection, video captioning and matching, and video search & retrieval. In *Proceedings of TRECVID*, volume 2019.

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.

Starkaður Barkarson and Steinþór Steingrímsson. 2019. Compiling and filtering ParIce: An English-Icelandic parallel corpus. In *Proceedings of the*

22nd Nordic Conference on Computational Linguistics, pages 140–145, Turku, Finland. Linköping University Electronic Press.

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*, 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*, Florence, Italy. Association for Computational Linguistics.

Alberto Barrón-Cedeño, Cristina España-Bonet, Josu Boldoba, and Lluís Màrquez. 2015. A Factory of Comparable Corpora from Wikipedia. In *Proceedings of the Eighth Workshop on Building and Using Comparable Corpora*, pages 3–13.

Chao Bei and Hao Zong. 2021. Gtcom neural machine translation systems for wmt21. In *Proceedings of the Sixth Conference on Machine Translation*, Online. 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, Matthias Huck, Philipp Koehn, 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*, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018a. 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.

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, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, 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*, 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, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors. 2019. *Proceedings of the Fourth Conference on Machine Translation*. Association for Computational Linguistics, Florence, Italy.

Ondřej Bojar, Jiří Mírovský, Kateřina Rysová, and Magdaléna Rysová. 2018b. EvalD Reference-Less Discourse Evaluation for WMT18. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Shaw 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 Shaw 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 10–48, Montreal, 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.

Sheila Castilho. 2020. On the same page? comparing inter-annotator agreement in sentence and document level human machine translation evaluation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1150–1159, Online. Association for Computational Linguistics.

Sheila Castilho, Maja Popović, and Andy Way. 2020. On context span needed for machine translation evaluation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3735–3742, Marseille, France. European Language Resources Association.

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 *Proc. of IWSLT*, Tokyo, Japan.

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)*, Beijing, China.

Pinzhen Chen, Jindřich Helcl, Ulrich Germann, Laurie Burchell, Nikolay Bogoychev, Antonio Valerio Miceli Barone, Jonas Waldendorf, Alexandra Birch, and Kenneth Heafield. 2021. The University of Edinburgh's English-German and English-Hausa submissions to the WMT21 news translation task. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Wei-Rui Chen and Muhammad Abdul-Mageed. 2021. Machine translation of low-resource indo-european languages. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Marta R. Costa-jussà, Marcos Zampieri, and Santanu Pal. 2018. A Neural Approach to Language Variety Translation. In *Proceedings of VarDial*.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9260–9269.

Rohit Dholakia and Anoop Sarkar. 2014. Pivot-based triangulation for low-resource languages. In *Proceedings of the Eleventh Conference of the Association for Machine Translation in the Americas (AMTA)*, volume 1, pages 315–328.

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.

Grant Erdmann, Jeremy Gwinnup, and Tim Anderson. 2021. Tune in: The afrl wmt21 news-translation systems. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Carlos Escolano, Ioannis Tsiamas, Christine Basta, Javier Ferrando, Marta R. Costa-jussà, and José A. R. Fonollosa. 2021. The talp-upc participation in wmt21 news translation task: an mbart-based nmt approach. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Cristina España-Bonet, Alberto Barrón-Cedeño, and Lluís Màrquez. 2020. Tailoring and Evaluating the Wikipedia for in-Domain Comparable Corpora Extraction.

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.

Christian Federmann. 2012. Appraise: an open-source toolkit for manual evaluation of mt output. *The Prague Bulletin of Mathematical Linguistics*, 98:25–35.

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.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

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 F. 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. *CoRR*, abs/2104.14478.

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*, Online. Association for Computational Linguistics.

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*, Online. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2021. The FLORES-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation. *arXiv:2106.03193 [cs]*. ArXiv: 2106.03193.

Yvette Graham, George Awad, and Alan Smeaton. 2018. Evaluation of automatic video captioning using direct assessment. *PLOS ONE*, 13(9):1–20.

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.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2016. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, pages 1–28.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2019. Translationese in Machine Translation Evaluation. *arXiv e-prints*, page arXiv:1906.09833.

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.

Roman Grundkiewicz, Marcin Junczys-Dowmunt, Christian Federmann, and Tom Kocmi. 2021. On user interfaces for large-scale document-level human evaluation of machine translation outputs. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 97–106, Online. Association for Computational Linguistics.

Hangcheng Guo, Wenbin Liu, Yanqing He, Tian Lan, Hongjiao Xu, Zhenfeng Wu, and You Pan. 2021. Istic's triangular machine translation system for wmt2021. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Barry Haddow and Faheem Kirefu. 2020. Pmindia–a collection of parallel corpora of languages of india. *arXiv preprint arXiv:2001.09907*.

Hossein Hassani. 2017. Kurdish Interdialect Machine Translation. *Proceedings of VarDial*.

Kenneth Heafield, Qianqian Zhu, and Roman Grundkiewicz. 2021. Findings of the wmt 2021 shared task on efficient translation. In *Proceedings of the*

Sixth Conference on Machine Translation, Online. Association for Computational Linguistics.

Amr Hendy, Esraa A. Gad, Mohamed Abdelghaffar, Jailan S. ElMosalami, Mohamed Afify, Ahmed Y. Tawfik, and Hany Hassan Awadalla. 2021. Ensembling of distilled models from multi-task teachers for constrained resource language pairs. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Josef Jon, Michal Novák, João Paulo Aires, Dusan Varis, and Ondřej Bojar. 2021. Cuni systems for wmt21: Multilingual low-resource translation for indo-european languages shared task. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Log-linear Combinations of Monolingual and Bilingual Neural Machine Translation Models for Automatic Post-Editing. In *Proceedings of the First Conference on Machine Translation*, pages 751–758, Berlin, Germany. Association for Computational Linguistics.

Haukur Jónsson, Haukur Barri Símonarson, Vésteinn Snæbjarnarson, Pétur Orri Ragnarson, and Vilhjálmur Þorsteinsson. 2021. Miðeind's wmt 2021 submission. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Ksenia Kharitonova, Ona de Gibert Bonet, Jordi Armengol-Estapé, Mar Rodriguez i Alvarez, and Maite Melero. 2021. Transfer learning with shallow decoders: Bsc at wmt2021's multilingual low-resource translation for indo-european languages shared task. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A. Smith, and Daniel S. Weld. 2021. GENIE: A leaderboard for human-in-the-loop evaluation of text generation.

Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based transfer learning for neural machine translation between non-english languages. *CoRR*, abs/1909.09524.

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. *arXiv e-prints*, page arXiv:2107.10821.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain.

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. 2006. Manual and Automatic Evaluation of Machine Translation between European Languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City. Association for Computational Linguistics.

Mikołaj Koszowski, Karol Grzegorczyk, and Tsimur Hadeliya. 2021. Allegro.eu submission to wmt21 news translation task. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Anoop Kunchukuttan. 2020. The IndicNLP Library. https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf.

Samuel Laubli, 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.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018a. 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.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018b. Has Neural Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In *EMNLP 2018*, Brussels, Belgium. Association for Computational Linguistics.

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*, Online. Association for Computational Linguistics.

Zongyao Li, Daimeng Wei, Hengchao Shang, Xiaoyu Chen, Zhanglin Wu, Zhengzhe Yu, Jiaxin Guo, Minghan Wang, Lizhi Lei, Min Zhang, Hao Yang, and Ying Qin. 2021a. Hw-tsc's participation in the wmt 2021 triangular mt shared task. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Zuchao Li, Masao Utiyama, Eiichiro Sumita, and Hai Zhao. 2021b. Miss@wmt21: Contrastive learning-reinforced domain adaptation in neural machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Jindřich Libovický and Alexander Fraser. 2021. Findings of the wmt 2021 shared tasks in unsupervised mt and very low resource supervised mt. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007.

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.

Huan Liu, Junpeng Liu, Kaiyu Huang, and Degen Huang. 2021a. Dutnlp machine translation system for wmt21 triangular translation task. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Pengfei Liu, Jinlan Fu, Yang Xiao, Weizhe Yuan, Shuaicheng Chang, Junqi Dai, Yixin Liu, Zihuiwen Ye, and Graham Neubig. 2021b. Explainaboard: An explainable leaderboard for nlp. *arXiv preprint arXiv:2104.06387*.

Shikun Liu, Edward Johns, and Andrew J. Davison. 2019. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Nikola Ljubešić and Antonio Toral. 2014. caWaC – a web corpus of Catalan and its application to language modeling and machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1728–1732, Reykjavik, Iceland. European Language Resources Association (ELRA).

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*, Brussels, Belgium. Association for Computational Linguistics.

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*, Online. Association for Computational Linguistics.

Ander Martinez. 2021. The fujitsu dmath submissions for wmt21 news translation and biomedical translation tasks. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Shivam Mhaskar and Pushpak Bhattacharyya. 2021. Pivot based transfer learning for neural machine translation: Cfilt iitb @ wmt 2021 triangular mt. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Simon Mille, Anja Belz, Bernd Bohnet, Yvette Graham, Emily Pitler, and Leo Wanner. 2018. The first multilingual surface realisation shared task (sr'18): Overview and evaluation results. In *Proceedings of the First Workshop on Multilingual Surface Realisation*, pages 1–12. Association for Computational Linguistics.

Simon Mille, Anja Belz, Bernd Bohnet, Yvette Graham, and Leo Wanner. 2019. The second multilingual surface realisation shared task (SR'19): Overview and evaluation results. In *Proceedings of the 2nd Workshop on Multilingual Surface Realisation (MSR 2019)*, pages 1–17, Hong Kong, China. 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 12th Language Resources and Evaluation Conference*, pages 3603–3609.

Matteo Negri, Marco Turchi, Rajen Chatterjee, and Nicola Bertoldi. 2018. eSCAPE: a Large-scale Synthetic Corpus for Automatic Post-Editing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. RankME: Reliable human ratings for natural language generation. 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 72–78, New Orleans, Louisiana. Association for Computational Linguistics.

Artur Nowakowski and Tomasz Dwojak. 2021. Adam mickiewicz university's english-hausa submissions to the wmt 2021 news translation task. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Shinhyeok Oh, Sion Jang, Hu Xu, Shounan An, and Insoo Oh. 2021. Netmarble AI Center's WMT21 Automatic Post-Editing Shared Task Submission. In *Proceedings of the Sixth Conference on Machine Translation*, Online.

Csaba Oravecz, Katina Bontcheva, David Kolovratník, Bhavani Bhaskar, Michael Jellinghaus, and Andreas Eisele. 2021. etranslation's submissions to the wmt 2021 news translation task. In *Proceedings of the Sixth Conference on Machine Translation*, Online. 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.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation (WMT)*.

Proyag Pal, Alham Fikri Aji, Pinzhen Chen, and Sukanta Sen. 2021. The University of Edinburgh's Bengali-Hindi submissions to the WMT21 news translation task. In *Proceedings of the Sixth Conference on Machine Translation*, 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 on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jeonghyeok Park, Hyunjoong Kim, and Hyunchang Cho. 2021. Papago's submissions to the wmt21 triangular translation task. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Nikita Pavlichenko, Ivan Stelmakh, and Dmitry Ustalov. 2021. Crowdspeech and vox DIY: Benchmark dataset for crowdsourced audio transcription. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Martin Popel and Ondřej Bojar. 2018. Training Tips for the Transformer Model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70.

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, Alberto Poncelas, Marija Brkic, and Andy Way. 2020. Neural machine translation for translating into Croatian and Serbian. In *Proceedings of the Seventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Lihua Qian, Hao Zhou, Yu Bao, Mingxuan Wang, Lin Qiu, Weinan Zhang, Yong Yu, and Lei Li. 2020. Glancing transformer for non-autoregressive neural machine translation. *arXiv preprint arXiv:2008.07905*.

Lihua Qian, Yi Zhou, Zaixiang Zheng, Yaoming ZHU, Zehui Lin, Jiangtao Feng, Shanbo Cheng, Lei Li, Mingxuan Wang, and Hao Zhou. 2021. The volctrans glat system: Non-autoregressive translation meets wmt21. In *Proceedings of the Sixth Conference on Machine Translation*, Online. 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.

Keisuke Sakaguchi and Benjamin Van Durme. 2018. Efficient online scalar annotation with bounded support. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 208–218, Melbourne, Australia. 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 e-prints*, page arXiv:1907.05791.

Abhishek Sharma, Prabhakar Gupta, and Anil Nelakanti. 2021. Adapting Neural Machine Translation for Automatic Post-Editing. In *Proceedings of the Sixth Conference on Machine Translation*, Online.

Shashank Siripragada, Jerin Philip, Vinay P. Namboodiri, and C V Jawahar. 2020. A multilingual parallel corpora collection effort for Indian languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3743–3751, Marseille, France. European Language Resources Association.

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.

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*, Online. Association for Computational Linguistics.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Steinþór Steingrímsson, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 2018. Risamálheild: A very large Icelandic text corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

Sandeep Subramanian, Oleksii Hrinchuk, Virginia Adams, and Oleksii Kuchaiev. 2021. Nvidia nemo's neural machine translation systems for english-german and english-russian news and biomedical tasks at wmt21. In *Proceedings of the Sixth Conference on Machine Translation*, 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.

Allahsera Auguste Tapo, Bakary Coulibaly, Sébastien Diarra, Christopher Homan, Julia Kreutzer, Sarah Luger, Arthur Nagashima, Marcos Zampieri, and Michael Leventhal. 2020. Neural machine translation for extremely low-resource african languages: A case study on bambara. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 23–32.

Svetlana Tchistiakova, Jesujoba Alabi, Koel Dutta Chowdhury, Sourav Dutta, and Dana Ruiter. 2021. Edinsaar@wmt21: North-germanic low-resource multilingual nmt. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Jörg Tiedemann. 2009. News from OPUS - a collection of multilingual parallel corpora with tools and interfaces. In *Proceedings of the Conference on Recent Advances in Natural Language Processing*, pages 237–248. John Benjamins.

Jörg Tiedemann and Lars Nygaard. 2004. The opus corpus-parallel and free: http://logos.uio.no/opus. In *Proceedings of LREC*.

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).

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018a. Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018b. 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, Belgium, Brussels. Association for Computational Linguistics.

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*, Online. Association for Computational Linguistics.

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.

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.

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.

Longyue Wang, Mu Li, Fangxu Liu, Shuming Shi, Zhaopeng Tu, Xing Wang, Shuangzhi Wu, Jiali Zeng, and Wen Zhang. 2021. Tencent translation system for the wmt21 news translation task. In *Proceedings of the Sixth Conference on Machine Translation*, 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, Hao Yang, and Ying Qin. 2021. Hw-tsc's participation in the wmt 2021 news translation shared task. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Guillaume Wenzek, Vishrav Chaudhary, Angela Fan, Sahir Gomez, Naman Goyal, Somya Jain, Douwe Kiela, Tristan Thrush, and Francisco Guzmán. 2021. Findings on the wmt 2021 shared task on large-scale multilingual machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Hua Wu and Haifeng Wang. 2009. Revisiting Pivot Language Approach for Machine Translation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International*

Joint Conference on Natural Language Processing of the AFNLP, pages 154–162, Suntec, Singapore. Association for Computational Linguistics.

Jitao Xu, Minh Quang Pham, Sadaf Abdul Rauf, and François Yvon. 2021. LISN @ WMT 2021. In *Proceedings of the Sixth Conference on Machine Translation*, Online. 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.

Han Yang, Bojie Hu, Wanying Xie, ambyera han, Pan Liu, Jinan Xu, and Qi Ju. 2021. Tentrans multilingual low-resource translation system for wmt21 indo-european languages task. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Lana Yeganova, Dina Wiemann, Mariana Neves, Federica Vezzani, Amy Siu, Inigo Jauregi Unanue, Maite Oronoz, Nancy Mah, Aurélie Névéol, David Martinez, Rachel Bawden, Giorgio Maria Di Nunzio, Roland Roller, Philippe Thomas, Cristian Grozea, Olatz Perez-de Viñaspre, Maika Vicente Navarro, and Antonio Jimeno Yepes. 2021. Findings of the wmt 2021 biomedical translation shared task: Summaries of animal experiments as new test set. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Hui Zeng. 2021. Small model and in-domain data are all you need. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Xianfeng Zeng, Yijin Liu, Ernan Li, Qiu Ran, Fandong Meng, Peng Li, Jinan Xu, and Jie Zhou. 2021. Wechat neural machine translation systems for wmt21. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

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, ACL 2020, Online, July 5-10, 2020*, pages 8545–8554. Association for Computational Linguistics.

Shiyu Zhao, Xiaopu Li, Minghui Wu, and Jie Hao. 2021. The mininglamp machine translation system for wmt21. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Shuhan Zhou, Tao Zhou, Binghao Wei, Yingfeng Luo, Yongyu Mu, Zefan Zhou, Chenglong Wang, Xuanjun Zhou, Chuanhao Lv, Yi Jing, Laohu Wang, Jingnan Zhang, Canan Huang, Zhongxiang Yan, Chi Hu, Bei Li, Tong Xiao, and Jingbo Zhu. 2021. The niutrans machine translation systems for wmt21. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

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.

A Differences in Human Scores

Tables 49–59 show differences in average standardized human scores for all pairs of competing systems for each language pair. The numbers in each of the tables' cells indicate the difference in average standardized 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 according to Wilcoxon rank-sum test ($p < 0.05$). Gray lines separate clusters based on non-overlapping rank ranges.

Tables 49-68 provide automatic metric scores (COMET, BLEU, chrF) for all competing systems.

	FACEBOOK-AI	ONLINE-A	CUNI-DOCTRANSFORMER	ONLINE-B	CUNI-TRANSFORMER2018	ONLINE-W	ONLINE-G	ONLINE-Y	HUMAN
FACEBOOK-AI	-	0.03	0.10⋆	0.12⋆	0.12‡	0.14‡	0.15‡	0.20‡	0.20‡
ONLINE-A	-0.03	-	0.07†	0.09†	0.09‡	0.11‡	0.12‡	0.17‡	0.17‡
CUNI-DOCTRANSFORMER	-0.10	-0.07	-	0.01	0.02	0.04	0.05†	0.09†	0.09⋆
ONLINE-B	-0.12	-0.09	-0.01	-	0.00	0.03	0.03⋆	0.08†	0.08†
CUNI-TRANSFORMER2018	-0.12	-0.09	-0.02	0.00	-	0.02	0.03	0.08⋆	0.08
ONLINE-W	-0.14	-0.11	-0.04	-0.03	-0.02	-	0.01	0.05⋆	0.05
ONLINE-G	-0.15	-0.12	-0.05	-0.03	-0.03	-0.01	-	0.05	0.05
ONLINE-Y	-0.20	-0.17	-0.09	-0.08	-0.08	-0.05	-0.05	-	0.00
HUMAN	-0.20	-0.17	-0.09	-0.08	-0.08	-0.05	-0.05	0.00	-
score	0.11	0.08	0.01	-0.01	-0.01	-0.03	-0.04	-0.08	-0.09
rank	1–2	1–2	3–6	3–6	3–8	3–8	5–9	7–9	5–9
bleu-A	31.1	28.3	30.2	**31.7**	26.2	28.9	28.6	24.6	-
chrF-A	**.599**	.569	.585	.593	.551	.576	.575	.549	-
comet-A	**.628**	.534	.592	.557	.510	.595	.517	.459	.358
bleu-B	**26.4**	23.5	24.7	24.8	21.7	24.8	22.8	20.3	-
chr-B	**.549**	.520	.532	.531	.504	.534	.520	.502	-
comet-B	**.513**	.411	.466	.431	.391	.486	.383	.322	.414

Table 49: Head to head comparison for Czech→English systems

	NiuTrans	KwaiNLP	DIDI-NLP	HUMAN	HappyNewYear	P3AI	Borderline	ICL	MiSS	IIE-MT	Machine-Translation	SMU	yyds	Facebook-AI	Online-B	ZengHuiMT	HuaweiTSC	XMU	NJUSC-TSC	Online-G	Online-W	UF	Online-A	Online-Y
NiuTrans	-	0.00	0.01	0.02	0.03	0.04★	0.07†	0.07†	0.07‡	0.07†	0.07‡	0.08‡	0.08‡	0.08‡	0.08‡	0.09‡	0.10‡	0.11‡	0.11‡	0.12‡	0.13‡	0.15‡	0.15‡	0.22‡
KwaiNLP	0.00	-	0.01	0.02	0.02	0.04	0.06★	0.06★	0.07†	0.07	0.07‡	0.07†	0.07†	0.08★	0.08†	0.09†	0.10‡	0.11‡	0.11‡	0.12‡	0.13‡	0.14‡	0.14‡	0.21‡
DIDI-NLP	-0.01	-0.01	-	0.01	0.02	0.03	0.05†	0.06†	0.06†	0.06★	0.06‡	0.07‡	0.07‡	0.07†	0.07‡	0.08‡	0.09‡	0.10‡	0.10‡	0.11‡	0.12‡	0.13‡	0.14‡	0.21‡
HUMAN	-0.02	-0.02	-0.01	-	0.00	0.02	0.04	0.04	0.05★	0.05	0.05†	0.05★	0.05★	0.06★	0.06★	0.07★	0.08†	0.09†	0.09‡	0.10‡	0.11†	0.12‡	0.13‡	0.19‡
HappyNewYear	-0.03	-0.02	-0.02	0.00	-	0.02	0.04	0.04	0.04★	0.05	0.05†	0.05★	0.05★	0.05★	0.06★	0.07★	0.08†	0.08†	0.09‡	0.10‡	0.10†	0.12‡	0.12‡	0.19‡
P3AI	-0.04	-0.04	-0.03	-0.02	-0.02	-	0.02	0.03	0.03	0.03	0.03†	0.03★	0.04★	0.04	0.04★	0.05★	0.06★	0.07★	0.07‡	0.08‡	0.09†	0.10‡	0.11‡	0.17‡
Borderline	-0.07	-0.06	-0.05	-0.04	-0.04	-0.02	-	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.03	0.04	0.05	0.05★	0.06★	0.06	0.08†	0.08‡	0.15‡
ICL	-0.07	-0.06	-0.06	-0.04	-0.04	-0.03	0.00	-	0.00	0.01	0.01★	0.01	0.01	0.01	0.02	0.02	0.04	0.04	0.05†	0.06†	0.06	0.08‡	0.08‡	0.15‡
MiSS	-0.07	-0.07	-0.06	-0.05	-0.04	-0.03	-0.01	0.00	-	0.00	0.00	0.01	0.01	0.01	0.01	0.02	0.03	0.04	0.04★	0.05★	0.06	0.07†	0.08‡	0.14†
IIE-MT	-0.07	-0.07	-0.06	-0.05	-0.05	-0.03	-0.01	-0.01	0.00	-	0.00★	0.00	0.00	0.01	0.01	0.02	0.03	0.04	0.04†	0.05†	0.06★	0.07†	0.08‡	0.14‡
Machine-Translation	-0.07	-0.07	-0.06	-0.05	-0.05	-0.03	-0.01	-0.01	0.00	0.00	-	0.00	0.00	0.01	0.01	0.02	0.03	0.04	0.04	0.05	0.05	0.07	0.07★	0.14★
SMU	-0.08	-0.07	-0.07	-0.05	-0.05	-0.03	-0.01	-0.01	-0.01	0.00	0.00	-	0.00	0.00	0.01	0.02	0.03	0.03	0.04	0.05	0.05	0.07★	0.07†	0.14†
yyds	-0.08	-0.07	-0.07	-0.05	-0.05	-0.04	-0.01	-0.01	-0.01	0.00	0.00	0.00	-	0.00	0.01	0.01	0.03	0.03	0.04	0.05	0.05	0.07★	0.07†	0.14★
Facebook-AI	-0.08	-0.08	-0.07	-0.06	-0.05	-0.04	-0.01	-0.01	-0.01	-0.01	-0.01	0.00	0.00	-	0.00	0.01	0.03	0.03	0.04★	0.04	0.05	0.07†	0.07†	0.14†
Online-B	-0.08	-0.08	-0.07	-0.06	-0.06	-0.04	-0.02	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01	0.00	-	0.01	0.02	0.03	0.03	0.04	0.04	0.06★	0.06†	0.13†
ZengHuiMT	-0.09	-0.09	-0.08	-0.07	-0.07	-0.05	-0.03	-0.02	-0.02	-0.02	-0.02	-0.02	-0.01	-0.01	-0.01	-	0.01	0.02	0.02	0.03	0.04	0.05★	0.06†	0.12†
HuaweiTSC	-0.10	-0.10	-0.09	-0.08	-0.08	-0.06	-0.04	-0.04	-0.03	-0.03	-0.03	-0.03	-0.03	-0.03	-0.02	-0.01	-	0.01	0.01	0.02	0.02	0.04★	0.04†	0.11★
XMU	-0.11	-0.11	-0.10	-0.09	-0.08	-0.07	-0.05	-0.04	-0.04	-0.04	-0.04	-0.03	-0.03	-0.03	-0.03	-0.02	-0.01	-	0.00	0.01	0.02	0.03	0.04†	0.11★
NJUSC-TSC	-0.11	-0.11	-0.10	-0.09	-0.09	-0.07	-0.05	-0.05	-0.04	-0.04	-0.04	-0.04	-0.04	-0.04	-0.03	-0.02	-0.01	0.00	-	0.01	0.01	0.03	0.03	0.10
Online-G	-0.12	-0.12	-0.11	-0.10	-0.10	-0.08	-0.06	-0.06	-0.05	-0.05	-0.05	-0.05	-0.05	-0.04	-0.04	-0.03	-0.02	-0.01	-0.01	-	0.01	0.02	0.02	0.09
Online-W	-0.13	-0.13	-0.12	-0.11	-0.10	-0.09	-0.06	-0.06	-0.06	-0.06	-0.05	-0.05	-0.05	-0.05	-0.04	-0.04	-0.02	-0.02	-0.01	-0.01	-	0.02	0.02★	0.09★
UF	-0.15	-0.14	-0.13	-0.12	-0.12	-0.10	-0.08	-0.08	-0.07	-0.07	-0.07	-0.07	-0.07	-0.07	-0.06	-0.05	-0.04	-0.03	-0.03	-0.02	-0.02	-	0.00	0.07
Online-A	-0.15	-0.14	-0.14	-0.13	-0.12	-0.11	-0.08	-0.08	-0.08	-0.08	-0.07	-0.07	-0.07	-0.07	-0.06	-0.06	-0.04	-0.04	-0.03	-0.02	-0.02	0.00	-	0.07
Online-Y	-0.22	-0.21	-0.21	-0.19	-0.19	-0.17	-0.15	-0.15	-0.14	-0.14	-0.14	-0.14	-0.14	-0.14	-0.13	-0.12	-0.11	-0.11	-0.10	-0.09	-0.09	-0.07	-0.07	-
score	0.04	0.04	0.03	0.02	0.02	-0.00	-0.02	-0.03	-0.03	-0.03	-0.03	-0.03	-0.04	-0.04	-0.04	-0.05	-0.06	-0.07	-0.07	-0.08	-0.09	-0.10	-0.11	-0.17
rank	1–5	1–6	1–6	1–9	1–9	2–19	4–18	4–19	6–17	3–24	9–22	7–21	7–21	6–20	7–21	7–21	7–21	7–22	12–24	11–24	8–22	17–24	20–24	20–24
bleu	31.9	32.2	32.5	-	29.3	29.6	**33.4**	30.1	32.6	31.7	29.8	32.2	32.7	32.1	32.4	32.2	28.9	31.9	28.7	27.2	28.5	30.2	24.6	24.1
chrF	.611	.611	.617	-	.584	.593	**.617**	.594	.613	.609	.588	.609	.612	.604	.613	.601	.593	.594	.590	.564	.575	.595	.545	.543
comet	**.487**	.476	.462	-	.417	.459	.483	.433	.483	.462	.433	.467	.478	.477	.438	.413	.434	.421	.427	.323	.358	.447	.253	.248

Table 50: Head to head comparison for Chinese→English systems

	BORDERLINE	ONLINE-A	ONLINE-W	UF	VOLCTRANS-AT	FACEBOOK-AI	ICL	ONLINE-G	ONLINE-B	ONLINE-Y	VOLCTRANS-GLAT	P3AI	SMU	UEDIN	NEMO	MANIFOLD	WATERMELON	HAPPYPOET	HUMAN	HUAWEITSC
BORDERLINE	-	0.00	0.00	0.01	0.02	0.03⋆	0.06⋆	0.08⋆	0.11‡	0.11‡	0.12†	0.12†	0.13‡	0.13‡	0.14‡	0.16‡	0.17‡	0.19‡	0.21‡	0.25‡
ONLINE-A	0.00	-	0.00	0.01	0.02	0.02	0.06⋆	0.08	0.11‡	0.11‡	0.11†	0.12†	0.13‡	0.13‡	0.13‡	0.16‡	0.17‡	0.19‡	0.21‡	0.24‡
ONLINE-W	0.00	0.00	-	0.01	0.02⋆	0.02⋆	0.05†	0.07†	0.11‡	0.11‡	0.11‡	0.11‡	0.13‡	0.13‡	0.13‡	0.16‡	0.16‡	0.18‡	0.20‡	0.24‡
UF	-0.01	-0.01	-0.01	-	0.01⋆	0.01⋆	0.05†	0.06⋆	0.10‡	0.10‡	0.10‡	0.11‡	0.12‡	0.12‡	0.12‡	0.15‡	0.16‡	0.17‡	0.19‡	0.23‡
VOLCTRANS-AT	-0.02	-0.02	-0.02	-0.01	-	0.01	0.04	0.06	0.09⋆	0.09⋆	0.10⋆	0.10	0.11†	0.11†	0.12†	0.14‡	0.15‡	0.17†	0.19‡	0.23‡
FACEBOOK-AI	-0.03	-0.02	-0.02	-0.01	-0.01	-	0.03	0.05	0.08†	0.08†	0.09⋆	0.09	0.11†	0.11†	0.11†	0.13‡	0.14‡	0.16⋆	0.18‡	0.22‡
ICL	-0.06	-0.06	-0.05	-0.05	-0.04	-0.03	-	0.02	0.05⋆	0.05	0.06	0.06	0.08⋆	0.08⋆	0.08⋆	0.10†	0.11‡	0.13	0.15‡	0.19‡
ONLINE-G	-0.08	-0.08	-0.07	-0.06	-0.06	-0.05	-0.02	-	0.03	0.03	0.04	0.04	0.06⋆	0.06⋆	0.06⋆	0.08†	0.09‡	0.11⋆	0.13‡	0.17‡
ONLINE-B	-0.11	-0.11	-0.11	-0.10	-0.09	-0.08	-0.05	-0.03	-	0.00	0.01	0.01	0.02	0.02	0.03	0.05	0.06⋆	0.08	0.10⋆	0.14‡
ONLINE-Y	-0.11	-0.11	-0.11	-0.10	-0.09	-0.08	-0.05	-0.03	0.00	-	0.01	0.01	0.02	0.02	0.03	0.05	0.06⋆	0.08	0.10†	0.14‡
VOLCTRANS-GLAT	-0.12	-0.11	-0.11	-0.10	-0.10	-0.09	-0.06	-0.04	-0.01	-0.01	-	0.00	0.02	0.02	0.02	0.04	0.05†	0.07	0.09†	0.13‡
P3AI	-0.12	-0.12	-0.11	-0.11	-0.10	-0.09	-0.06	-0.04	-0.01	-0.01	0.00	-	0.01	0.02	0.02	0.04⋆	0.05†	0.07	0.09†	0.13‡
SMU	-0.13	-0.13	-0.13	-0.12	-0.11	-0.11	-0.08	-0.06	-0.02	-0.02	-0.02	-0.01	-	0.00	0.00	0.03	0.03	0.05	0.07	0.11†
UEDIN	-0.13	-0.13	-0.13	-0.12	-0.11	-0.11	-0.08	-0.06	-0.02	-0.02	-0.02	-0.02	0.00	-	0.00	0.03	0.03⋆	0.05	0.07⋆	0.11†
NEMO	-0.14	-0.13	-0.13	-0.12	-0.12	-0.11	-0.08	-0.06	-0.03	-0.03	-0.02	-0.02	0.00	0.00	-	0.02	0.03⋆	0.05	0.07⋆	0.11‡
MANIFOLD	-0.16	-0.16	-0.16	-0.15	-0.14	-0.13	-0.10	-0.08	-0.05	-0.05	-0.04	-0.04	-0.03	-0.03	-0.02	-	0.01	0.03	0.05	0.08⋆
WATERMELON	-0.17	-0.17	-0.16	-0.16	-0.15	-0.14	-0.11	-0.09	-0.06	-0.06	-0.05	-0.05	-0.03	-0.03	-0.03	-0.01	-	0.02	0.04	0.08
HAPPYPOET	-0.19	-0.19	-0.18	-0.17	-0.17	-0.16	-0.13	-0.11	-0.08	-0.08	-0.07	-0.07	-0.05	-0.05	-0.05	-0.03	-0.02⋆	-	0.02⋆	0.06‡
HUMAN	-0.21	-0.21	-0.20	-0.19	-0.19	-0.18	-0.15	-0.13	-0.10	-0.10	-0.09	-0.09	-0.07	-0.07	-0.07	-0.05	-0.04	-0.02	-	0.04
HUAWEITSC	-0.25	-0.24	-0.24	-0.23	-0.23	-0.22	-0.19	-0.17	-0.14	-0.14	-0.13	-0.13	-0.11	-0.11	-0.11	-0.08	-0.08	-0.06	-0.04	-
score	0.13	0.12	0.12	0.11	0.11	0.10	0.07	0.05	0.02	0.02	0.01	0.01	-0.01	-0.01	-0.01	-0.04	-0.04	-0.06	-0.08	-0.12
rank	1–5	1–6	1–4	4	3–8	4–9	5–12	4–12	8–17	7–17	7–17	5–16	9–19	9–17	9–17	10–19	15–20	7–17	16–20	18–20
bleu-A	34.9	34.0	32.9	33.8	34.4	34.0	32.5	33.4	33.8	32.1	**35.0**	33.2	33.8	33.7	33.3	33.0	34.5	31.2	-	34.6
chrF-A	**.629**	.625	.619	.619	.623	.625	.616	.619	.619	.605	.625	.614	.622	.620	.617	.614	.623	.592	-	.625
comet-A	.612	.616	.622	.618	.607	.638	.618	.593	.617	.556	.612	.600	.589	.608	.621	.590	.608	.450	.531	**.628**
bleu-B	39.0	39.6	39.7	40.1	39.7	**41.9**	39.6	38.8	40.2	37.5	39.5	37.9	38.7	40.2	40.6	37.8	39.8	34.6	-	40.8
chrF-B	.654	.657	.661	.658	.652	**.671**	.660	.649	.657	.634	.652	.643	.651	.656	.656	.642	.652	.610	-	.659
comet-B	.620	.628	.650	.638	.626	**.670**	.642	.609	.638	.568	.624	.612	.599	.638	.641	.605	.626	.446	.564	.644

Table 51: Head to head comparison for German→English systems

	NEMO	ONLINE-W	ONLINE-B	HUMAN	MANIFOLD	FACEBOOK-AI	NIUTRANS	ONLINE-G	AFRL	ONLINE-A	ONLINE-Y
NEMO	-	0.01	0.03⋆	0.05	0.08	0.08	0.09⋆	0.12‡	0.15‡	0.16‡	0.26‡
ONLINE-W	-0.01	-	0.02⋆	0.04	0.07⋆	0.07	0.09†	0.11‡	0.14‡	0.15‡	0.25‡
ONLINE-B	-0.03	-0.02	-	0.02	0.05	0.05	0.06	0.09†	0.12‡	0.13‡	0.23‡
HUMAN	-0.05	-0.04	-0.02	-	0.03	0.03	0.04	0.07†	0.10‡	0.11‡	0.21‡
MANIFOLD	-0.08	-0.07	-0.05	-0.03	-	0.00	0.02	0.04⋆	0.07†	0.08†	0.18‡
FACEBOOK-AI	-0.08	-0.07	-0.05	-0.03	0.00	-	0.01	0.04†	0.07‡	0.08‡	0.18‡
NIUTRANS	-0.09	-0.09	-0.06	-0.04	-0.02	-0.01	-	0.03	0.06⋆	0.07⋆	0.17‡
ONLINE-G	-0.12	-0.11	-0.09	-0.07	-0.04	-0.04	-0.03	-	0.03	0.04	0.14⋆
AFRL	-0.15	-0.14	-0.12	-0.10	-0.07	-0.07	-0.06	-0.03	-	0.01	0.11
ONLINE-A	-0.16	-0.15	-0.13	-0.11	-0.08	-0.08	-0.07	-0.04	-0.01	-	0.10
ONLINE-Y	-0.26	-0.25	-0.23	-0.21	-0.18	-0.18	-0.17	-0.14	-0.11	-0.10	-
score	0.14	0.13	0.11	0.09	0.06	0.06	0.04	0.02	-0.01	-0.02	-0.12
rank	1–5	1–4	3–7	1–7	2–7	1–7	3–8	7-10	8–11	8–11	9–11
bleu-A	40.2	37.0	40.6	-	41.1	**42.3**	41.8	41.2	38.8	38.7	32.8
chrF-A	.660	.631	.661	-	.659	.661	.658	**.668**	.635	.652	.600
comet-A	.625	.610	.624	.619	.619	**.656**	.632	.635	.595	.595	.524
bleu-B	40.1	37.2	40.0	-	40.5	**41.6**	41.2	40.7	39.6	38.8	33.2
chrF-B	.663	.635	.663	-	.661	.663	.661	**.671**	.640	.657	.602
comet-B	.619	.606	.621	.619	.614	**.647**	.623	.629	.589	.591	.523

Table 52: Head to head comparison for Russian→English systems

	HUAWEITSC	IIE-MT	NIUTRANS	KWAINLP	FACEBOOK-AI	XMU	CAPITALMARVEL	ONLINE-B	MISS	ONLINE-W	WECHAT-AI	ONLINE-A	ONLINE-G	MOVELIKEAJAGUAR	ONLINE-Y	ILLINI
HUAWEITSC	-	0.06⋆	0.09⋆	0.11‡	0.11⋆	0.12‡	0.13‡	0.14‡	0.17‡	0.18‡	0.20‡	0.22‡	0.28‡	0.30‡	0.33‡	0.33‡
IIE-MT	-0.06	-	0.04	0.05	0.05	0.06†	0.07⋆	0.08†	0.11†	0.12†	0.14‡	0.16‡	0.22‡	0.24‡	0.27‡	0.27‡
NIUTRANS	-0.09	-0.04	-	0.01	0.01	0.02⋆	0.03	0.04⋆	0.08†	0.08⋆	0.11‡	0.13‡	0.19‡	0.20‡	0.23‡	0.24‡
KWAINLP	-0.11	-0.05	-0.01	-	0.00	0.01	0.02	0.03	0.06⋆	0.07	0.10†	0.11‡	0.17‡	0.19‡	0.22‡	0.23‡
FACEBOOK-AI	-0.11	-0.05	-0.01	0.00	-	0.01⋆	0.02	0.03⋆	0.06⋆	0.07⋆	0.09‡	0.11‡	0.17‡	0.19‡	0.22‡	0.22‡
XMU	-0.12	-0.06	-0.02	-0.01	-0.01	-	0.01	0.02	0.06	0.06	0.09	0.11⋆	0.17‡	0.18‡	0.21‡	0.22‡
CAPITALMARVEL	-0.13	-0.07	-0.03	-0.02	-0.02	-0.01	-	0.01	0.04	0.05	0.07†	0.09‡	0.15‡	0.17‡	0.20‡	0.20‡
ONLINE-B	-0.14	-0.08	-0.04	-0.03	-0.03	-0.02	-0.01	-	0.03	0.04	0.06	0.08⋆	0.14‡	0.16‡	0.19‡	0.19‡
MISS	-0.17	-0.11	-0.08	-0.06	-0.06	-0.06	-0.04	-0.03	-	0.01	0.03	0.05⋆	0.11‡	0.13†	0.16‡	0.16‡
ONLINE-W	-0.18	-0.12	-0.08	-0.07	-0.07	-0.06	-0.05	-0.04	-0.01	-	0.02	0.04†	0.10‡	0.12‡	0.15‡	0.15‡
WECHAT-AI	-0.20	-0.14	-0.11	-0.10	-0.09	-0.09	-0.07	-0.06	-0.03	-0.02	-	0.02	0.08⋆	0.09⋆	0.13†	0.13†
ONLINE-A	-0.22	-0.16	-0.13	-0.11	-0.11	-0.11	-0.09	-0.08	-0.05	-0.04	-0.02	-	0.06	0.08	0.11⋆	0.11⋆
ONLINE-G	-0.28	-0.22	-0.19	-0.17	-0.17	-0.17	-0.15	-0.14	-0.11	-0.10	-0.08	-0.06	-	0.02	0.05	0.05
MOVELIKEAJAGUAR	-0.30	-0.24	-0.20	-0.19	-0.19	-0.18	-0.17	-0.16	-0.13	-0.12	-0.09	-0.08	-0.02	-	0.03	0.04
ONLINE-Y	-0.33	-0.27	-0.23	-0.22	-0.22	-0.21	-0.20	-0.19	-0.16	-0.15	-0.13	-0.11	-0.05	-0.03	-	0.00
ILLINI	-0.33	-0.27	-0.24	-0.23	-0.22	-0.22	-0.20	-0.19	-0.16	-0.15	-0.13	-0.11	-0.05	-0.04	0.00	-
score	0.14	0.08	0.05	0.03	0.03	0.03	0.01	0.00	-0.03	-0.04	-0.06	-0.08	-0.14	-0.16	-0.19	-0.19
rank	1	2–5	2–6	2–9	2–6	5–11	3–10	5–11	6–11	5–11	7–12	11–14	12–16	12–16	13–16	13–16
bleu	26.5	25.4	27.2	25.8	27.7	25.8	23.7	27.2	27.0	22.8	**27.8**	21.0	20.6	21.2	17.3	18.6
chrf	.528	.521	.532	.524	**.536**	.524	.496	.526	.529	.489	.535	.455	.476	.476	.482	.453
comet	.348	.314	.371	.307	**.392**	.307	.236	.270	.294	.270	.361	.167	.145	.182	.061	.073

Table 53: Head to head comparison for Japanese→English systems

	FACEBOOK-AI	MANIFOLD	NIUTRANS	ONLINE-B	HUAWEITSC	MIDEIND	ONLINE-A	ALLEGRO	ONLINE-Y	ONLINE-G
FACEBOOK-AI	-	0.18‡	0.25‡	0.26‡	0.28‡	0.28‡	0.29‡	0.33‡	0.37‡	0.55‡
MANIFOLD	-0.18	-	0.07⋆	0.08‡	0.10†	0.10†	0.11‡	0.15‡	0.19‡	0.37‡
NIUTRANS	-0.25	-0.07	-	0.02	0.03	0.04	0.04	0.08†	0.12†	0.30‡
ONLINE-B	-0.26	-0.08	-0.02	-	0.02	0.02	0.03	0.07	0.11⋆	0.28‡
HUAWEITSC	-0.28	-0.10	-0.03	-0.02	-	0.00	0.01	0.05⋆	0.09†	0.27‡
MIDEIND	-0.28	-0.10	-0.04	-0.02	0.00	-	0.01	0.05⋆	0.09⋆	0.26‡
ONLINE-A	-0.29	-0.11	-0.04	-0.03	-0.01	-0.01	-	0.04	0.08	0.26‡
ALLEGRO	-0.33	-0.15	-0.08	-0.07	-0.05	-0.05	-0.04	-	0.04	0.22‡
ONLINE-Y	-0.37	-0.19	-0.12	-0.11	-0.09	-0.09	-0.08	-0.04	-	0.18‡
ONLINE-G	-0.55	-0.37	-0.30	-0.28	-0.27	-0.26	-0.26	-0.22	-0.18	-
score	0.29	0.11	0.04	0.03	0.01	0.01	0.00	-0.04	-0.08	-0.26
rank	1	2	3–7	3–8	3–7	3–7	3–9	6–9	7–9	10
bleu	**41.7**	39.8	39.2	40.6	38.4	33.5	33.6	33.3	30.1	23.7
chrF	.623	.621	.610	**.624**	.611	.578	.574	.574	.559	.492
comet	**.683**	.629	.619	.645	.604	.552	.512	.467	.422	-.071

Table 54: Head to head comparison for Icelandic→English systems

	FACEBOOK-AI	ONLINE-B	TRANSSION	ZMT	GTCOM	HUAWEITSC	MS-EGDC	P3AI	NIUTRANS	ONLINE-Y	MANIFOLD	AMU	UEDIN	TWB
FACEBOOK-AI	-	0.13‡	0.19‡	0.19‡	0.19‡	0.22‡	0.25‡	0.28‡	0.28‡	0.34‡	0.36‡	0.42‡	0.45‡	0.51‡
ONLINE-B	-0.13	-	0.06⋆	0.06	0.06	0.09†	0.12‡	0.15‡	0.15‡	0.21‡	0.23‡	0.29‡	0.32‡	0.39‡
TRANSSION	-0.19	-0.06	-	0.00	0.00	0.03	0.06	0.09†	0.09⋆	0.15‡	0.17‡	0.24‡	0.27‡	0.33‡
ZMT	-0.19	-0.06	0.00	-	0.00	0.03	0.06⋆	0.09†	0.09⋆	0.15‡	0.17‡	0.23‡	0.26‡	0.33‡
GTCOM	-0.19	-0.06	0.00	0.00	-	0.03	0.06⋆	0.09†	0.09⋆	0.15‡	0.17‡	0.23‡	0.26‡	0.33‡
HUAWEITSC	-0.22	-0.09	-0.03	-0.03	-0.03	-	0.03	0.06	0.06	0.12†	0.14‡	0.20‡	0.23‡	0.30‡
MS-EGDC	-0.25	-0.12	-0.06	-0.06	-0.06	-0.03	-	0.03	0.03	0.09⋆	0.11†	0.18‡	0.21‡	0.27‡
P3AI	-0.28	-0.15	-0.09	-0.09	-0.09	-0.06	-0.03	-	0.00	0.06	0.08⋆	0.14†	0.17‡	0.24‡
NIUTRANS	-0.28	-0.15	-0.09	-0.09	-0.09	-0.06	-0.03	0.00	-	0.06	0.08⋆	0.14‡	0.17‡	0.24‡
ONLINE-Y	-0.34	-0.21	-0.15	-0.15	-0.15	-0.12	-0.09	-0.06	-0.06	-	0.02	0.08⋆	0.12†	0.18‡
MANIFOLD	-0.36	-0.23	-0.17	-0.17	-0.17	-0.14	-0.11	-0.08	-0.08	-0.02	-	0.06	0.09⋆	0.16‡
AMU	-0.42	-0.29	-0.24	-0.23	-0.23	-0.20	-0.18	-0.14	-0.14	-0.08	-0.06	-	0.03	0.09†
UEDIN	-0.45	-0.32	-0.27	-0.26	-0.26	-0.23	-0.21	-0.17	-0.17	-0.12	-0.09	-0.03	-	0.06⋆
TWB	-0.51	-0.39	-0.33	-0.33	-0.33	-0.30	-0.27	-0.24	-0.24	-0.18	-0.16	-0.09	-0.06	-
score	0.25	0.12	0.06	0.06	0.06	0.03	0.00	-0.03	-0.03	-0.09	-0.11	-0.17	-0.20	-0.27
rank	1	2–4	3–7	2–6	3–6	3–9	5–19	6–10	6–10	8–11	10–12	11–13	12–13	14
bleu	**21.0**	18.7	18.8	18.8	17.8	17.5	17.1	17.8	16.5	13.9	16.9	14.1	14.9	12.3
chrF	**.487**	.467	.472	.472	.467	.468	.453	.463	.447	.448	.456	.413	.422	.403
comet	**.422**	.335	.345	.344	.345	.253	.148	.245	.174	.124	.127	.070	.076	-0.046

Table 55: Head to head comparison for Hausa→English systems

	GTCOM	ONLINE-B	TRANSSION	MS-EGDC	UEDIN	ONLINE-Y	HUAWEITSC	ONLINE-A	ONLINE-G
GTCOM	-	0.04	0.12‡	0.13‡	0.15‡	0.22‡	0.28‡	0.31‡	0.58‡
ONLINE-B	-0.04	-	0.08⋆	0.09⋆	0.11†	0.18‡	0.24‡	0.27‡	0.54‡
TRANSSION	-0.12	-0.08	-	0.00	0.03	0.09⋆	0.16†	0.19†	0.45‡
MS-EGDC	-0.13	-0.09	0.00	-	0.02	0.09	0.16†	0.18†	0.45‡
UEDIN	-0.15	-0.11	-0.03	-0.02	-	0.07	0.13⋆	0.16†	0.43‡
ONLINE-Y	-0.22	-0.18	-0.09	-0.09	-0.07	-	0.07	0.09	0.36‡
HUAWEITSC	-0.28	-0.24	-0.16	-0.16	-0.13	-0.07	-	0.03	0.29‡
ONLINE-A	-0.31	-0.27	-0.19	-0.18	-0.16	-0.09	-0.03	-	0.27‡
ONLINE-G	-0.58	-0.54	-0.45	-0.45	-0.43	-0.36	-0.29	-0.27	-
score	0.20	0.16	0.08	0.08	0.05	-0.01	-0.08	-0.11	-0.37
rank	1–2	1–2	3–5	3–5	3–6	4–8	6–8	6–8	9
bleu	24.2	24.1	**24.5**	21.1	21.7	21.5	21.9	21.1	16.7
chrF	**.517**	.512	.512	.486	.489	.488	.488	.483	.433
comet	**.692**	.670	.637	.532	.584	.501	.528	.494	.116

Table 56: Head to head comparison for Bengali→Hindi systems

	HuaweiTSC	Online-A	GTCOM	UEdin	Online-Y	TRANSSION	Online-B	MS-EgDC	Online-G
HuaweiTSC	-	0.01	0.01	0.03	0.17⋆	0.20‡	0.22⋆	0.25‡	1.35‡
Online-A	-0.01	-	0.00	0.02	0.16⋆	0.19‡	0.21⋆	0.24‡	1.34‡
GTCOM	-0.01	0.00	-	0.02	0.15†	0.19‡	0.20†	0.24‡	1.33‡
UEdin	-0.03	-0.02	-0.02	-	0.13⋆	0.17‡	0.19⋆	0.22‡	1.31‡
Online-Y	-0.17	-0.16	-0.15	-0.13	-	0.04⋆	0.05	0.09‡	1.18‡
TRANSSION	-0.20	-0.19	-0.19	-0.17	-0.04	-	0.02	0.05⋆	1.14‡
Online-B	-0.22	-0.21	-0.20	-0.19	-0.05	-0.02⋆	-	0.04†	1.13‡
MS-EgDC	-0.25	-0.24	-0.24	-0.22	-0.09	-0.05	-0.04	-	1.09‡
Online-G	-1.35	-1.34	-1.33	-1.31	-1.18	-1.14	-1.13	-1.09	-
score	0.24	0.24	0.23	0.21	0.08	0.04	0.03	-0.01	-1.10
rank	1–4	1–4	1–4	1–4	5–6	7	6–7	8	9
bleu	13.0	13.4	13.9	12.5	10.6	15.0	**15.3**	10.9	5.9
chrF	.457	.465	.471	.454	.432	.478	**.480**	.434	.364
comet	.523	.552	**.575**	.545	.386	.537	.535	.411	-0.215

Table 57: Head to head comparison for Hindi→Bengali systems

	TRANSSION	HuaweiTSC	MS-EgDC	GTCOM	Online-G
TRANSSION	-	0.19‡	0.24‡	0.34‡	1.75‡
HuaweiTSC	-0.19	-	0.05	0.15†	1.56‡
MS-EgDC	-0.24	-0.05	-	0.10	1.51‡
GTCOM	-0.34	-0.15	-0.10	-	1.41‡
Online-G	-1.75	-1.56	-1.51	-1.41	-
score	0.50	0.31	0.26	0.16	-1.25
rank	1	2–3	2–4	3–4	5
bleu	**14.5**	9.9	9.2	11.9	3.6
chrF	**.503**	.486	.476	.475	.361
comet	.290	**.315**	.299	.199	-.606

Table 58: Head to head comparison for Zulu→Xhosa systems

	HuaweiTSC	TRANSSION	GTCOM	MS-EgDC	FJDMATH	Online-G
HuaweiTSC	-	0.04	0.09	0.19‡	0.22‡	1.47‡
TRANSSION	-0.04	-	0.05	0.14†	0.18‡	1.42‡
GTCOM	-0.09	-0.05	-	0.10⋆	0.13†	1.38‡
MS-EgDC	-0.19	-0.14	-0.10	-	0.04	1.28‡
FJDMATH	-0.22	-0.18	-0.13	-0.04	-	1.24‡
Online-G	-1.47	-1.42	-1.38	-1.28	-1.24	-
score	0.33	0.29	0.24	0.14	0.11	-1.14
rank	1–3	1–3	1–3	4–5	4–5	6
bleu	**11.8**	**11.8**	11.5	9.9	9.8	3.9
chrF	**.504**	.497	.493	.477	.479	.370
comet	**.233**	.206	.192	.180	.197	-.582

Table 59: Head to head comparison for Xhosa→Zulu systems

Rank	Ave.	Ave. z	System	Comet_A	$\text{BLEU}_{A,B}$	BLEU_A	BLEU_B	chrF_A	chrF_B
1	90.2	0.397	HUMAN-A	–	–	–	–	–	–
2-4	87.9	0.284	HUMAN-B	–	–	–	–	–	–
2-4	87.6	0.263	Facebook-AI	**0.775**	**36.1**	**24.8**	**22.7**	**0.536**	**0.506**
2-4	86.1	0.214	Online-W	0.751	33.6	23.0	21.6	0.528	0.500
5-7	83.0	0.122	eTranslation	0.625	30.8	21.0	19.4	0.506	0.478
5-6	82.1	0.047	CUNI-Transformer2018	0.671	31.5	21.6	19.7	0.509	0.482
6-8	79.2	-0.120	CUNI-DocTransformer	0.680	32.1	22.2	19.8	0.517	0.485
7-9	79.3	-0.154	CUNI-Marian-Baselines	0.621	28.9	20.1	18.3	0.499	0.472
8-10	77.8	-0.183	Online-B	0.586	28.9	20.0	17.9	0.496	0.466
9-10	74.6	-0.308	Online-A	0.585	29.0	20.2	18.2	0.499	0.468
11	76.2	-0.373	Online-Y	0.456	26.2	18.1	16.1	0.481	0.451
12	65.6	-0.674	Online-G	0.293	22.0	15.3	13.9	0.457	0.431

Table 60: Automatic metric scores for English→Czech systems

Rank	Ave.	Ave. z	System	Comet_A	Comet_C	$\text{BLEU}_{A,C}$	BLEU_A	BLEU_C	chrF_A	chrF_C
1-17	83.3	0.266	Online-B	0.502	0.568	47.3	28.4	37.2	0.588	0.650
1-5	84.7	0.243	Online-W	0.546	0.616	51.0	29.7	41.3	0.602	0.678
1-14	86.6	0.217	WeChat-AI	0.548	0.610	**51.2**	**31.3**	40.0	**0.607**	0.668
1-6	87.6	0.145	Facebook-AI	**0.567**	**0.630**	52.5	31.3	**42.0**	0.606	**0.676**
1-10	89.4	0.116	UF	0.507	0.573	47.3	28.5	37.2	0.589	0.650
2-17	85.2	0.089	HW-TSC	0.516	0.576	48.9	29.8	38.6	0.597	0.658
3-17	86.8	0.072	UEdin	0.517	0.574	48.4	29.9	38.0	0.595	0.650
3-18	86.5	0.041	P3AI	0.498	0.560	46.3	28.3	36.5	0.584	0.639
3-18	86.4	0.030	HUMAN-A	–	0.554	–	–	–	–	–
5-19	83.3	0.013	happypoet	0.452	0.511	44.6	27.6	35.4	0.582	0.634
4-19	86.1	0.010	eTranslation	0.506	0.568	48.7	29.6	38.5	0.594	0.653
4-19	84.4	0.001	Online-A	0.511	0.573	47.6	29.0	37.9	0.594	0.653
3-18	84.5	0.001	HUMAN-C	0.540	–	–	–	–	–	–
5-19	78.8	-0.053	VolcTrans-AT	0.518	0.580	47.8	29.3	38.0	0.595	0.653
5-19	86.7	-0.055	NVIDIA-NeMo	0.531	0.592	49.8	30.0	39.2	0.598	0.660
8-21	83.1	-0.058	Manifold	0.497	0.557	47.5	29.4	37.2	0.592	0.644
4-20	84.3	-0.062	Online-G	0.439	0.497	43.4	27.1	33.5	0.577	0.627
12-20	84.5	-0.072	Online-Y	0.465	0.522	45.2	27.9	35.3	0.582	0.636
18-21	73.9	-0.130	ICL	0.196	0.246	39.0	24.5	30.4	0.552	0.595
4-20	85.0	-0.140	VolcTrans-GLAT	0.542	0.616	53.6	31.3	43.2	0.608	0.683
16-21	78.3	-0.179	nuclear_trans	0.386	0.445	44.3	27.7	34.5	0.578	0.626
22	80.0	-0.415	BUPT_rush	0.371	0.428	42.0	26.4	32.6	0.571	0.618

Table 61: Automatic metric scores for English→German systems

Rank	Ave.	Ave. z	System	Comet_A	BLEU_A	chrF_A
1-2	84.1	0.362	HUMAN-A	–	–	–
1-4	82.7	0.264	Facebook-AI	**0.329**	20.1	0.511
2-5	80.8	0.263	NiuTrans	0.304	19.7	**0.532**
3-6	81.2	0.175	Online-B	0.224	18.9	0.504
4-6	80.1	0.128	TRANSSION	0.228	18.9	0.504
2-6	79.2	0.124	ZMT	0.230	18.8	0.504
7-10	78.0	0.018	P3AI	0.273	**20.4**	0.517
7-10	78.7	0.006	HW-TSC	0.307	20.3	0.512
8-12	75.2	-0.026	AMU	0.092	16.2	0.465
7-10	78.8	-0.036	GTCOM	0.197	17.9	0.499
9-12	75.0	-0.128	MS-EgDC	0.086	16.1	0.465
12-15	70.2	-0.227	UEdin	-0.061	14.8	0.453
11-15	73.4	-0.243	Manifold	0.175	18.0	0.495
12-15	70.5	-0.340	TWB	0.000	17.1	0.483
11-15	67.7	-0.448	Online-Y	0.083	15.0	0.469

Table 62: Automatic metric scores for English→Hausa systems

Rank	Ave.	Ave. z	System	Comet$_A$	BLEU$_A$	chrF$_A$
1	88.1	0.872	HUMAN-A	–	–	–
2	84.5	0.594	Facebook-AI	**0.776**	**33.3**	**0.596**
3-4	68.2	0.277	NiuTrans	0.694	30.6	0.575
3-4	72.7	0.240	Manifold	0.648	28.6	0.562
5-9	75.2	0.200	Online-A	0.550	25.5	0.545
5-7	65.6	0.130	Lan-Bridge-MT	0.589	24.9	0.538
5-9	62.6	0.063	Mideind	0.542	24.3	0.531
6-9	73.9	0.026	Online-B	0.583	25.7	0.543
6-9	75.6	-0.034	HW-TSC	0.560	27.5	0.554
10	62.0	-0.236	Online-Y	0.351	22.4	0.513
11	48.7	-0.470	Allegro.eu	0.323	22.7	0.510
12	33.9	-1.082	Online-G	-0.327	12.2	0.421

Table 63: Automatic metric scores for English→Icelandic systems

Rank	Ave.	Ave. z	System	Comet$_A$	BLEU$_A$	chrF$_A$
1-2	86.4	0.430	Facebook-AI	**0.652**	46.8	0.407
1-2	85.3	0.314	HUMAN-A	–	–	–
3-5	84.2	0.266	Online-W	0.602	42.1	0.366
3-5	81.3	0.168	WeChat-AI	0.615	**46.9**	**0.404**
3-5	82.6	0.148	NiuTrans	0.619	46.2	0.399
6-8	77.8	0.017	HW-TSC	0.614	45.4	0.392
6-8	71.8	-0.042	MiSS	0.517	42.6	0.370
8-13	78.5	-0.051	Online-Y	0.386	39.5	0.341
6-10	77.8	-0.067	BUPT_rush	0.549	42.9	0.372
8-13	70.9	-0.129	Online-A	0.421	40.8	0.350
9-13	67.4	-0.184	Online-B	0.488	41.6	0.360
9-14	74.2	-0.284	ephemeraler	0.414	39.6	0.343
9-14	72.5	-0.339	capitalmarvel	0.460	41.0	0.355
12-14	70.1	-0.373	movelikeajaguar	0.379	38.5	0.334
15-16	63.5	-0.440	Illini	0.189	34.3	0.294
15-16	65.7	-0.541	Online-G	0.143	33.5	0.287

Table 64: Automatic metric scores for English→Japanese systems

Rank	Ave.	Ave. z	System	Comet$_A$	Comet$_B$	BLEU$_{A,B}$	BLEU$_A$	BLEU$_B$	chrF$_A$	chrF$_B$
1-3	86.0	0.317	HUMAN-B	0.600	–	–	–	–	–	–
1-3	83.3	0.277	Online-W	**0.664**	**0.660**	45.0	31.8	29.9	0.576	**0.571**
1-3	82.5	0.093	HUMAN-A	–	0.599	–	–	–	–	–
4-6	79.4	0.056	Online-B	0.604	0.601	43.5	29.8	29.2	0.568	0.567
4-7	75.3	0.032	Online-A	0.576	0.559	41.2	28.8	27.2	0.561	0.556
4-7	80.1	-0.001	Facebook-AI	0.650	0.644	**46.0**	**32.2**	**30.4**	**0.576**	0.571
7-10	74.5	-0.123	NiuTrans	0.512	0.510	40.5	28.4	27.1	0.546	0.543
7-10	72.3	-0.153	Manifold	0.566	0.566	41.5	29.2	27.6	0.554	0.551
7-10	75.4	-0.161	NVIDIA-NeMo	0.582	0.578	41.6	29.3	27.6	0.562	0.558
5-10	76.0	-0.180	Online-G	0.600	0.595	42.8	30.1	28.6	0.570	0.564
11	62.7	-0.541	Online-Y	0.474	0.470	37.7	25.8	25.3	0.538	0.538

Table 65: Automatic metric scores for English→Russian systems

Rank	Ave.	Ave. z	System	Comet$_A$	Comet$_B$	BLEU$_{A,B}$	BLEU$_A$	BLEU$_B$	chrF$_A$	chrF$_B$
1-3	82.5	0.325	HUMAN-B	0.427	–	–	–	–	–	–
2-14	74.9	0.284	HappyNewYear	0.468	0.403	48.0	35.7	32.1	0.300	0.278
1-7	81.2	0.250	Facebook-AI	0.499	0.425	49.9	35.9	35.3	0.343	0.331
1-8	80.0	0.216	HUMAN-A	–	0.421	–	–	–	–	–
4-19	75.3	0.164	Borderline	0.473	0.403	49.2	36.5	33.2	0.313	0.289
2-19	81.0	0.161	bjtu_nmt	0.474	0.409	46.9	34.8	32.5	0.295	0.274
3-14	75.5	0.151	Lan-Bridge-MT	0.463	0.406	44.6	32.6	31.3	0.320	0.300
4-21	79.3	0.124	BUPT_rush	0.425	0.368	44.7	33.1	31.1	0.296	0.278
2-18	79.2	0.098	NiuTrans	0.483	0.411	48.1	35.8	32.9	0.305	0.282
4-18	75.7	0.091	Machine_Translation	0.467	0.403	47.7	35.5	32.3	0.294	0.275
2-15	80.9	0.078	SMU	0.474	0.402	47.9	35.8	32.5	0.306	0.280
6-22	81.4	0.064	capitalmarvel	0.378	0.299	43.9	32.2	30.5	0.268	0.261
4-19	79.5	0.056	WeChat-AI	0.501	0.437	49.2	36.9	33.4	0.337	0.305
6-22	78.1	0.026	Online-W	0.468	0.391	44.8	33.4	30.9	0.303	0.277
7-22	75.2	0.004	ICL	0.463	0.396	47.5	34.8	33.3	0.317	0.300
9-23	75.9	-0.008	HW-TSC	0.447	0.380	47.4	35.1	32.3	0.298	0.279
5-23	78.2	-0.025	ZengHuiMT	0.448	0.386	48.5	35.9	32.6	0.304	0.282
11-22	81.2	-0.026	yyds	0.474	0.407	48.1	35.9	32.4	0.302	0.278
10-26	79.7	-0.050	P3AI	0.436	0.375	47.0	34.0	33.3	0.318	0.308
17-27	77.1	-0.061	windfall	0.395	0.313	44.2	32.6	30.3	0.282	0.269
6-24	78.9	-0.075	Online-B	0.458	0.381	48.5	36.0	33.1	0.321	0.299
13-26	76.8	-0.080	NJUSC_TSC	0.439	0.381	46.3	34.2	31.9	0.312	0.291
9-24	77.7	-0.100	MiSS	0.468	0.404	49.0	36.2	33.2	0.304	0.286
19-27	77.0	-0.101	UF	0.413	0.361	45.3	33.1	31.4	0.288	0.277
22-28	72.7	-0.123	Online-A	0.340	0.292	43.3	31.6	30.1	0.264	0.261
22-28	79.3	-0.160	happypoet	0.364	0.307	43.5	32.5	29.7	0.277	0.259
20-28	76.9	-0.185	nuclear_trans	0.428	0.361	44.7	33.4	30.5	0.284	0.261
25-29	76.4	-0.247	ephemeraler	0.382	0.311	44.0	32.6	30.2	0.287	0.273
28-31	67.5	-0.257	Online-G	0.301	0.238	43.2	31.1	29.7	0.304	0.288
29-31	67.1	-0.463	Online-Y	0.317	0.254	43.9	32.0	30.9	0.281	0.271
29-31	68.3	-0.613	movelikeajaguar	0.371	0.309	43.7	32.7	29.7	0.280	0.260

Table 66: Automatic metric scores for English→Chinese systems

Rank	Ave.	Ave. z	System	Comet$_A$	BLEU$_A$	chrF$_A$
1–5	87.7	0.088	Online-W	**0.714**	**60.4**	**0.788**
1–7	89.2	0.052	Online-A	0.566	40.6	0.670
1–4	89.5	0.035	HUMAN-A	– –	–	
2–8	85.7	0.002	LISN	0.505	37.3	0.644
1–8	86.9	-0.014	Online-B	0.576	43.8	0.689
4–10	85.0	-0.021	talp_upc	0.481	36.3	0.641
3–8	85.0	-0.064	eTranslation	0.595	40.6	0.666
7–10	84.1	-0.154	Online-G	0.454	36.9	0.653
3–10	86.6	-0.210	Online-Y	0.503	39.5	0.659
7–10	86.4	-0.229	P3AI	0.583	39.3	0.654

Table 67: Automatic metric scores for French→German systems

Rank	Ave.	Ave. z	System	Comet$_A$	BLEU$_A$	chrF$_A$
1–3	87.9	0.160	Online-B	0.544	29.7	0.584
1–3	86.5	0.126	HUMAN-A	–	–	–
3–6	83.4	0.018	Manifold	0.586	32.5	0.606
1–6	84.8	0.006	Online-W	**0.622**	29.9	0.591
3–6	84.5	0.004	Online-A	0.561	**35.7**	0.613
6–10	83.0	-0.084	Online-G	0.449	28.6	0.577
3–10	83.5	-0.148	P3AI	0.512	31.7	**0.626**
6–10	81.3	-0.149	LISN	0.426	28.1	0.563
6–10	83.7	-0.177	Online-Y	0.463	28.3	0.568
6–10	81.0	-0.190	talp_upc	0.466	27.5	0.565

Table 68: Automatic metric scores for German→French systems

B Translator Brief: Sentence-Split News Test Sets

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.
- 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.

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.

C News Task System Submission Summaries

This appendix lists self-reported details on MT systems participating in the News Translation Task.

C.1 AFRL (Erdmann et al., 2021)

No brief description provided.

C.2 ALLEGRO.EU (Koszowski et al., 2021)

Allegro news translation system is based on the transformer-big architecture, it makes use of corpora filtering and backtranslation both applied to parallel and monolingual data alike.

ALLEGRO.EU	common	Multilingual MT System: No. Basic System Classification: Seq2seq Transformer Style [Vaswani+2017] (self-attention, ...) Token Unit Type Used: Unigram (as in https://github.com/google/sentencepiece) Vocabulary Size: 32000 Toolkit Used: OpenNMT-py Batch size: 8192 tokens Features of your model structure: Dropout, Tied source and target word embeddings Document-level training: No document-level: Our system processes each segment independently. Number of GPUs Used Concurrently: 1x A100 Wallclock training time: 13h Number of contrastive configurations used: 4 Other comments: fp16 was used
ALLEGRO.EU	en-is	True Parallel Training Data Size in Sentence Pairs: 3935903 parallel.en-is True Parallel Training Data Size in Words: 60185218 parallel.en 55419088 parallel.is Synthetic Parallel Training Data Size in Sentence Pairs: 2953528 synt.en-is Synthetic Parallel Training Data Size in Words: 47082741 synt.en 44441374 synt.is Monolingual Training Data in Sentences: 4044137 mono.en-is Monolingual Training Data in Words: 81559107 mono.en 72315845 mono.is Processing Tools Used: Language detection (e.g. for data cleanup) Features of your model development: Data filtering, Data selection, Iterative back-translation, Oversampling Number of Systems Ensembled/Averaged: 1
ALLEGRO.EU	is-en	True Parallel Training Data Size in Sentence Pairs: 3935903 parallel.is-en True Parallel Training Data Size in Words: 55419088 parallel.is 60185218 parallel.en Synthetic Parallel Training Data Size in Sentence Pairs: 2907611 synt.is-en Synthetic Parallel Training Data Size in Words: 43642048 synt.is 47392565 synt.en Monolingual Training Data in Sentences: 3991420 mono.is-en Monolingual Training Data in Words: 78481284 mono.is 81693347 mono.en Processing Tools Used: Tokenizer, Language detection (e.g. for data cleanup) Features of your model development: Data filtering, Data selection, Iterative back-translation, Oversampling, Ensembling Number of Systems Ensembled/Averaged: 2

C.3 AMU (Nowakowski and Dwojak, 2021)

AMU submission for the low-resource English-Hausa language pair involved data filtering and cleaning, transfer learning from the pretrained unrelated high-resource language pair (German-English) and iterative backtranslation. The initial iteration of backtranslation was performed with a PB-SMT model, while the subsequent iterations were performed with NMT Transformer models.

C.4 BJTU-NMT (no associated paper)

No brief description provided.

C.5 BORDERLINE (Wang et al., 2021)

No brief description provided.

C.6 BUPT-RUSH (no associated paper)

No brief description provided.

C.7 CAPITALMARVEL (no associated paper)

No brief description provided.

C.8 CFILT

We train our DE-DSB system using transfer learning from DE-HSB model. Our DE-HSB model is using monolingual data of HSB and DE and train an unsupervised system first using MASS objective, then finetune it with iterative back-translation and then finetune it for translation using parallel data of DE-HSB. This system is then trained using monolingual data of DE and DSB with iterative back-translation. We use shared encoder and decoder with 6 layers in both encoder and decoder.

CFILT	common	Multilingual MT System: No.
CFILT	de-dsb	Basic System Classification: Masked sequence to sequence pretraining (Song et al 2019)+ Transfer learning Token Unit Type Used: BPE (as in https://github.com/rsennrich/subword-nmt), Moses Tokenizer Vocabulary Size: 33678 True Parallel Training Data Size in Sentence Pairs: de-hsb 147521 de-dsb 0 Processing Tools Used: Tokenizer Other Processing Tools Used: fastBPE Toolkit Used: Moses, fastBPE, MASS Features of your model development: Iterative back-translation, Unsupervised (i.e. not involving parallel data), Language model pretraining with MASS objective Pre-trained parts of models: Masked Sequence to Sequence Pre-training (MASS) Document-level training: No document-level: Our system processes each segment independently. Other Features of Your Training: Transfer learning
CFILT	de-hsb	Basic System Classification: MASS pretraining (song et al) Token Unit Type Used: Unigram (as in https://github.com/google/sentencepiece), Moses Tokenizer Toolkit Used: Moses, fastBPE, MASS Pre-trained parts of models: Masked Sequence to Sequence Pre-training (MASS) Document-level training: No document-level: Our system processes each segment independently.
CFILT	dsb-de	Basic System Classification: MASS pretraining, Transfer learning Token Unit Type Used: BPE (as in https://github.com/rsennrich/subword-nmt), Moses Tokenizer
CFILT	hsb-de	Basic System Classification: MASS pretraining (song et al 2019), Transfer learning Token Unit Type Used: BPE (as in https://github.com/rsennrich/subword-nmt) Pre-trained parts of models: Masked Sequence to Sequence Pre-training (MASS)

C.9 CUNI (Gebauer et al., 2021)

CUNI-DOCTRANSFORMER CUNI-DocTransformer is similar to the sentence-level version called CUBBITT (Popel et al., 2020), but trained on sequences with multiple sentences of up to 3000 characters. This year, a better sentence detection and number/unit conversion post-processing have been applied.

CUNI-TRANSFORMER2018 CUNI-Transformer2018, also called CUBBITT, is exactly the same system as in WMT2018. It is the Transformer model trained according to Popel and Bojar (2018) plus a Block Back-translation (Popel et al., 2020).

CUNI	common	Multilingual MT System: No. Basic System Classification: Seq2seq Transformer Style [Vaswani+2017] (self-attention, ...) Token Unit Type Used: SubwordTextEncoder of Tensor2Tensor (as in https://github.com/tensorflow/tensor2tensor) Vocabulary Size: 32k Monolingual Training Data in Sentences: see synthetic Monolingual Training Data in Words: see synthetic Processing Tools Used: Tokenizer Toolkit Used: Tensor2Tensor Features of your model development: Data filtering, Data selection, Block-backtranslation as in Martin Popel, Marketa Tomkova, Jakub Tomek et al. (2020), Iterative back-translation, Oversampling, Averaging Features of your model structure: Dropout, Tied source and target word embeddings, Weight tying (other than word embeddings) Number of Systems Ensembled/Averaged: 8 checkpoints Wallclock training time: 8 days (without iterated backtranslation)
CUNI-DocTransformer	cs-en, en-cs	True Parallel Training Data Size in Sentence Pairs: 61000000 True Parallel Training Data Size in Words: en=617000000, cs=702000000 Synthetic Parallel Training Data Size in Sentence Pairs: en=76000000, cs=51000000 Synthetic Parallel Training Data Size in Words: en=1296000000, cs=833000000 Batch size: 1800*10 subwords Document-level training: Overlapping windows: A window is moved over segments, receiving multiple translations of each of them, with some voting or combination afterwards. Number of GPUs Used Concurrently: 10 GTX 1080 Ti Number of contrastive configurations used: 4
CUNI-Transformer2018	cs-en, en-cs	True Parallel Training Data Size in Sentence Pairs: 58000000 True Parallel Training Data Size in Words: en=642000000, cs=563000000 Synthetic Parallel Training Data Size in Sentence Pairs: en=47000000, cs=65000000 Synthetic Parallel Training Data Size in Words: en=935000000, cs=927000000 Batch size: 2900*8 subwords Document-level training: No document-level: Our system processes each segment independently. Number of GPUs Used Concurrently: 8 GTX 1080 Ti Number of contrastive configurations used: Now only one. In 2018, I trained hundreds of models on smaller data or less GPUs, as described in Training Tips for the Transformer Model (Popel and Bojar, 2018).

C.10 DIDI-NLP (no associated paper)

No brief description provided.

C.11 EPHEMERALER

We use Transformer big model and ensembling.

EPHEMERALER	common	Multilingual MT System: No. Basic System Classification: Seq2seq Transformer Style [Vaswani+2017] (self-attention, ...)
EPHEMERALER	en-ja	Token Unit Type Used: BPE (as in https://github.com/rsennrich/subword-nmt)
EPHEMERALER	en-zh	—

C.12 eTranslation (Oravecz et al., 2021)

eTranslations's En-De system is an ensemble of 4 big transformers, trained from all available parallel data (cleaned up and filtered with heuristic rules and with a language model built from the German NewsCrawl data) and with additional tagged, back-translated data generated from the monolingual news corpora. The original parallel data is upsampled to a 1:1 ratio. Each transformer model is then tuned on a 10M top subset of original parallel data scored and ranked by the monolingual news language model and then fine-tuned further on previous year's test sets. The models use a 36k SentencePiece vocabulary. The SentencePiece module as built in the Marian toolkit is used for end-to-end text processing, without the standard pre- and postprocessing steps of truecasing, or (de)tokenization.

The Fr-De system is an ensemble of 4 big transformers. Three of them are trained on original parallel (OP) data and back-translated (BT) data in a 1:1 ratio. The 4th big transformer was additionally finetuned for 7 epochs on 2M of the OP data scored by a domain language model. BT data and data for the domain language model were selected using topic modelling techniques to tune the model towards the domain defined in the task.

The En-Cs system is an ensemble of two big transformer models from last year's submission, trained on the WMT 2020 data, both original parallel and back-translated. Training on the 2021 data had not finished until the submission deadline and intermediate models scored worse than the 2020 models.

eTranslation	common	Multilingual MT System: No. Basic System Classification: Seq2seq Transformer Style [Vaswani+2017] (self-attention, ...) Token Unit Type Used: Unigram (as in https://github.com/google/sentencepiece) Toolkit Used: Marian Document-level training: No document-level: Our system processes each segment independently.
eTranslation	en-de	Vocabulary Size: 36000 True Parallel Training Data Size in Sentence Pairs: 32077088 True Parallel Training Data Size in Words: 637753194; 603406453 Synthetic Parallel Training Data Size in Sentence Pairs: 226375233 Synthetic Parallel Training Data Size in Words: 3514437534; 3007895939 Monolingual Training Data in Sentences: BT: 226375233; En LM: 133385694; De LM: 167110102; Monolingual Training Data in Words: BT: 3514437534; 3007895939 En LM: 2891767899; De LM: 3012152905 Processing Tools Used: Tokenizer, Language detection (e.g. for data cleanup) Batch size: 1500-5000 Features of your model development: Data filtering, Data selection, Back-translation with greedy decoding, Oversampling, Ensembling, Fine-tuning for domain adaptation Features of your model structure: Dropout, Tied source and target word embeddings Other Features of Your Training: continued training on LM scored subset of OP data Number of Systems Ensembled/Averaged: 4 Number of GPUs Used Concurrently: 4-8 V100 Wallclock training time: 10 days Number of contrastive configurations used: 16 Other comments: described in the system paper
eTranslation	fr-de	Vocabulary Size: 30000 True Parallel Training Data Size in Sentence Pairs: 13640043 True Parallel Training Data Size in Words: 257966051; 228953683 Synthetic Parallel Training Data Size in Sentence Pairs: 14980793 Synthetic Parallel Training Data Size in Words: 241457887; 209714902 Monolingual Training Data in Sentences: de: 11475958 Monolingual Training Data in Words: de: 160803597 Processing Tools Used: Tokenizer, Language detection (e.g. for data cleanup) Batch size: 1500 Features of your model development: Data filtering, Data selection, Back-translation with greedy decoding, Oversampling, Ensembling, Fine-tuning for domain adaptation Features of your model structure: Dropout, Tied source and target word embeddings Number of Systems Ensembled/Averaged: 4 Number of GPUs Used Concurrently: 4 Wallclock training time: 5 days Number of contrastive configurations used: 11
eTranslation	en-cs	Vocabulary Size: 36000 True Parallel Training Data Size in Sentence Pairs: 45104433 True Parallel Training Data Size in Words: cs: 559485115 en: 637004843 Synthetic Parallel Training Data Size in Sentence Pairs: 88164502 Synthetic Parallel Training Data Size in Words: cs: 1206604906 en: 1450464754 Monolingual Training Data in Sentences: 0 Monolingual Training Data in Words: 0 Processing Tools Used: Language detection (e.g. for data cleanup) Batch size: 1000 Features of your model development: Data filtering, Back-translation with sampling, Ensembling Features of your model structure: Dropout Number of Systems Ensembled/Averaged: 2 Number of GPUs Used Concurrently: 4 Wallclock training time: 12 days Number of contrastive configurations used: 4

C.13 FACEBOOK-AI (Tran et al., 2021)

Facebook AI participated in the unconstrained track for all 14 English-centric directions. To explore the limit of scaling multilingual translation, we trained two multilingual systems: Any-to-English, and English-to-Any, and submitted them to all directions. In addition to well-known techniques such as large scale backtranslation, in-domain finetuning, ensembling, and noisy channel re-ranking, we also experimented with scaling dense transformer (up to 4.7B parameters), and sparse mixture of experts (up to 52B parameters)

FACEBOOK-AI	common	Multilingual MT System: Yes, the system was trained and used jointly for all the language pairs. Basic System Classification: Seq2seq Transformer Style [Vaswani+2017] (self-attention, ...) Token Unit Type Used: BPE (as in https://github.com/rsennrich/subword-nmt)
FACEBOOK-AI	cs-en, de-en, ha-en, is-en, ja-en, ru-en, zh-en	Vocabulary Size: 128000 True Parallel Training Data Size in Sentence Pairs: (This includes mined data from CCMatrix and CCAligned) cs-en 163,005,937 de-en 544,549,887 ha-en 1,176,367 is-en 20,632,971 ja-en 141,399,044 ru-en 276,805,988 zh-en 163,188,501 Total 1,310,758,695 True Parallel Training Data Size in Words: (This includes mined data from CCMatrix and CCAligned) 2725979073 train.cs_en.cs 2661179726 train.cs_en.en 10546303763 train.de_en.de 9692849751 train.de_en.en 20466571 train.ha_en.ha 18786730 train.ha_en.en 342802801 train.is_en.is 301337746 train.is_en.en 640041697 train.ja_en.ja 1907474016 train.ja_en.en 4896618898 train.ru_en.ru 4887514242 train.ru_en.en 714086693 train.zh_en.zh 2853757236 train.zh_en.en Synthetic Parallel Training Data Size in Sentence Pairs: (Backtranslation data) cs-en 428,914,158 de-en 394,678,147 ha-en 378,439,788 is-en 428,581,678 ja-en 428,227,231 ru-en 381,863,501 zh-en 432,017,983 Total 2,872,722,486 Monolingual Training Data in Sentences: Similar to backtranslation data (430M English sentences) Processing Tools Used: Language detection (e.g. for data cleanup) Toolkit Used: fairseq(-py) Batch size: 1M tokens Features of your model development: Data filtering, Iterative back-translation, Ensembling, Averaging, Right-to-left reranking, Target-to-source reranking, Fine-tuning for domain adaptation, Mixture of Experts Features of your model structure: Dropout, Tied source and target word embeddings Document-level training: No document-level: Our system processes each segment independently. Other Features of Your Training: In-domain parallel data mining Number of Systems Ensembled/Averaged: 3 Number of GPUs Used Concurrently: 128 Wallclock training time: 1 week Number of contrastive configurations used: 5 different architectures, 3-4 training iterations each
FACEBOOK-AI	en-cs, en-de, en-ha, en-is, en-ja, en-ru, en-zh	Vocabulary Size: 128000 True Parallel Training Data Size in Sentence Pairs: (Includes mined data from CCMatrix, CCAligned) en-cs 163,758,080 en-de 546,657,024 en-ha 995,860 en-is 27,228,288 en-ja 142,843,968 en-ru 277,540,224 en-zh 163,774,144 Total 1,322,797,588 Synthetic Parallel Training Data Size in Sentence Pairs: en-cs 140,172,928 en-de 237,235,904 en-ha 6,719,488 en-is 101,139,008 en-ja 218,456,960 en-ru 163,223,744 en-zh 123,211,776 Total 990,159,808 Monolingual Training Data in Sentences: Same as backtranslation Processing Tools Used: Language detection (e.g. for data cleanup) Toolkit Used: fairseq(-py) Batch size: 1M tokens per batch Features of your model development: Data filtering, Data selection, Iterative back-translation, Oversampling, Ensembling, Averaging, Right-to-left reranking, Target-to-source reranking, Fine-tuning for domain adaptation Features of your model structure: Dropout, Tied source and target word embeddings Document-level training: No document-level: Our system processes each segment independently. Number of Systems Ensembled/Averaged: 2-3 Number of GPUs Used Concurrently: 128 Wallclock training time: 1 week Number of contrastive configurations used: 20

C.14 FJDMATH (Martinez, 2021)

No brief description provided.

C.15 GTCOM (Bei and Zong, 2021)

No brief description provided.

C.16 HAPPYNEWYEAR (no associated paper)

No brief description provided.

C.17 HAPPYPOET (no associated paper)

No brief description provided.

C.18 HW-TSC (Wei et al., 2021)

We participate in 7 language pairs including Zh/En, De/En, Ja/En, Ha/En, Is/En, Hi/Bn, and Xh/Zu and in both directions under the constrained condition. We use the standard Transformer-Big model as the baseline and obtain the best performance via two variants with larger parameter sizes. We perform detailed pre-processing and filtering on the provided large-scale bilingual and monolingual datasets. Several commonly used strategies are used to train our models such as Back Translation, Ensemble Knowledge Distillation, etc. We also conduct experiments regarding similar language augmentation, which lead to positive results, although not used in our submission. Our submission obtains competitive results in the final evaluation.

HW-TSC	common	Basic System Classification: Seq2seq Transformer Style [Vaswani+2017] (self-attention, ...) Document-level training: No document-level: Our system processes each segment independently. Number of GPUs Used Concurrently: 8
HW-TSC	en-zh	Multilingual MT System: No. Token Unit Type Used: BPE (as in https://github.com/rsennrich/subword-nmt), Moses Tokenizer, jieba Vocabulary Size: 32k True Parallel Training Data Size in Sentence Pairs: 16.5M Synthetic Parallel Training Data Size in Sentence Pairs: 316.5M Monolingual Training Data in Sentences: 300M Processing Tools Used: Tokenizer, Word Aligner (e.g. fast_align or GIZA++), Language detection (e.g. for data cleanup), Jieba word segmentation for Chinese Toolkit Used: Marian, fairseq(-py), Moses Batch size: 4096 Features of your model development: Data filtering, Data selection, Back-translation with sampling, Iterative back-translation, Forward translation for synthetic data, Ensembling, Averaging, Fine-tuning for domain adaptation Features of your model structure: Dropout Number of Systems Ensembled/Averaged: 2Ensembled
HW-TSC	zh-en	Multilingual MT System: No. Token Unit Type Used: BPE (as in https://github.com/rsennrich/subword-nmt), Moses Tokenizer, jieba Vocabulary Size: 32k True Parallel Training Data Size in Sentence Pairs: 16.5M Synthetic Parallel Training Data Size in Sentence Pairs: 316.5M Monolingual Training Data in Sentences: 300M Processing Tools Used: Tokenizer, Language detection (e.g. for data cleanup) Toolkit Used: Marian, fairseq(-py), Moses Batch size: 4096 Features of your model development: Data filtering, Data selection, Back-translation with sampling, Iterative back-translation, Forward translation for synthetic data, Ensembling, Averaging Features of your model structure: Dropout Number of Systems Ensembled/Averaged: 2ensemble

HW-TSC	en-ha	Multilingual MT System: Yes, the system was trained and used jointly for all the language pairs. Token Unit Type Used: Unigram (as in https://github.com/google/sentencepiece) Vocabulary Size: 32K True Parallel Training Data Size in Sentence Pairs: 0.6M Synthetic Parallel Training Data Size in Sentence Pairs: 14.9M Monolingual Training Data in Sentences: 14.3M Processing Tools Used: Word Aligner (e.g. fast_align or GIZA++), Language detection (e.g. for data cleanup) Toolkit Used: Marian, fairseq(-py) Features of your model development: Data filtering, Data selection, Back-translation with greedy decoding, Iterative back-translation, Forward translation for synthetic data, Ensembling, Averaging, Fine-tuning for domain adaptation Features of your model structure: Dropout Number of Systems Ensembled/Averaged: 4ensemble
HW-TSC	ha-en	Multilingual MT System: Yes, the system was trained and used jointly for all the language pairs. Vocabulary Size: 32K True Parallel Training Data Size in Sentence Pairs: 0.6M Synthetic Parallel Training Data Size in Sentence Pairs: 14.9M Monolingual Training Data in Sentences: 14.3M Processing Tools Used: Word Aligner (e.g. fast_align or GIZA++), Language detection (e.g. for data cleanup) Toolkit Used: Marian, fairseq(-py) Features of your model development: Data filtering, Data selection, Back-translation with greedy decoding, Iterative back-translation, Ensembling, Averaging, Fine-tuning for domain adaptation Features of your model structure: Dropout Number of Systems Ensembled/Averaged: 4
HW-TSC	en-is	Multilingual MT System: Yes, the system was trained and used jointly for all the language pairs. Token Unit Type Used: Unigram (as in https://github.com/google/sentencepiece) Vocabulary Size: 32K True Parallel Training Data Size in Sentence Pairs: 4M Synthetic Parallel Training Data Size in Sentence Pairs: 42M Monolingual Training Data in Sentences: 38M Processing Tools Used: Word Aligner (e.g. fast_align or GIZA++), Language detection (e.g. for data cleanup) Toolkit Used: Marian, fairseq(-py) Batch size: 4096 Features of your model development: Data filtering, Data selection, Knowledge distillation, Back-translation with greedy decoding, Iterative back-translation, Forward translation for synthetic data, Ensembling, Averaging, Fine-tuning for domain adaptation Features of your model structure: Dropout Number of Systems Ensembled/Averaged: 3
HW-TSC	is-en	Multilingual MT System: Yes, the system was trained and used jointly for all the language pairs. Token Unit Type Used: Unigram (as in https://github.com/google/sentencepiece) Vocabulary Size: 32K True Parallel Training Data Size in Sentence Pairs: 4M Synthetic Parallel Training Data Size in Sentence Pairs: 42M Monolingual Training Data in Sentences: 38M Processing Tools Used: Word Aligner (e.g. fast_align or GIZA++), Language detection (e.g. for data cleanup) Toolkit Used: Marian, fairseq(-py) Features of your model development: Data filtering, Data selection, Back-translation with greedy decoding, Iterative back-translation, Forward translation for synthetic data, Ensembling, Averaging, Fine-tuning for domain adaptation Features of your model structure: Dropout Number of Systems Ensembled/Averaged: 3

HW-TSC	bn-hi	Multilingual MT System: Yes, the system was trained and used jointly for all the language pairs. Token Unit Type Used: sentencepiece Vocabulary Size: 32000 True Parallel Training Data Size in Sentence Pairs: 3400000 Synthetic Parallel Training Data Size in Sentence Pairs: 46500000 Monolingual Training Data in Sentences: 46500000 Monolingual Training Data in Words: 1899414973 Processing Tools Used: Tokenizer, Word Aligner (e.g. fast_align or GIZA++), Language detection (e.g. for data cleanup) Toolkit Used: Marian, fairseq(-py) Batch size: 1500 Features of your model development: Data filtering, Data selection, Knowledge distillation, Back-translation with sampling, Iterative back-translation, Forward translation for synthetic data, Oversampling Number of Systems Ensembled/Averaged: 4
HW-TSC	hi-bn	Multilingual MT System: Yes, the system was trained and used jointly for all the language pairs. Token Unit Type Used: sentencepiece Vocabulary Size: 32000 True Parallel Training Data Size in Sentence Pairs: 3400000 Synthetic Parallel Training Data Size in Sentence Pairs: 50000000 Monolingual Training Data in Sentences: 50000000 Processing Tools Used: Tokenizer, Word Aligner (e.g. fast_align or GIZA++), Language detection (e.g. for data cleanup) Toolkit Used: Marian, fairseq(-py) Batch size: 1500 Features of your model development: Data filtering, Data selection, Knowledge distillation, Back-translation with sampling, Iterative back-translation, Forward translation for synthetic data, Oversampling, Ensembling, Averaging Number of Systems Ensembled/Averaged: 4
HW-TSC	xh-zu	Multilingual MT System: Yes, the system was trained and used jointly for all the language pairs. Token Unit Type Used: sentencepiece Vocabulary Size: 32000 True Parallel Training Data Size in Sentence Pairs: 67000 Synthetic Parallel Training Data Size in Sentence Pairs: 12000000 Monolingual Training Data in Sentences: 12000000 Processing Tools Used: Tokenizer, Word Aligner (e.g. fast_align or GIZA++), Language detection (e.g. for data cleanup) Toolkit Used: Marian, fairseq(-py) Batch size: 1500 Features of your model development: Data filtering, Data selection, Knowledge distillation, Back-translation with sampling, Iterative back-translation, Forward translation for synthetic data, Oversampling, Ensembling, Averaging, Fine-tuning for domain adaptation Number of Systems Ensembled/Averaged: 4
HW-TSC	zu-xh	Multilingual MT System: Yes, the system was trained and used jointly for all the language pairs. Token Unit Type Used: sentencepiece Vocabulary Size: 32000 True Parallel Training Data Size in Sentence Pairs: 67000 Synthetic Parallel Training Data Size in Sentence Pairs: 12000000 Synthetic Parallel Training Data Size in Words: 50000000 Processing Tools Used: Tokenizer, Word Aligner (e.g. fast_align or GIZA++) Toolkit Used: Marian, fairseq(-py) Batch size: 1500 Features of your model development: Data filtering, Data selection, Knowledge distillation, Back-translation with sampling, Iterative back-translation, Forward translation for synthetic data, Oversampling, Ensembling, Averaging Number of Systems Ensembled/Averaged: 4

HW-TSC	en-ja	Multilingual MT System: No. Token Unit Type Used: sentencepiece Vocabulary Size: 32000 True Parallel Training Data Size in Sentence Pairs: 14000000 Synthetic Parallel Training Data Size in Sentence Pairs: 80000000 Monolingual Training Data in Sentences: 150000000 Processing Tools Used: Tokenizer, Word Aligner (e.g. fast_align or GIZA++), Language detection (e.g. for data cleanup) Toolkit Used: Marian, fairseq(-py) Batch size: 1500 Features of your model development: Data filtering, Data selection, Knowledge distillation, Back-translation with sampling, Iterative back-translation, Forward translation for synthetic data, Oversampling, Ensembling, Averaging, Fine-tuning for domain adaptation Number of Systems Ensembled/Averaged: 4
HW-TSC	ja-en	Multilingual MT System: No. Token Unit Type Used: sentencepiece Vocabulary Size: 32000 True Parallel Training Data Size in Sentence Pairs: 12000000 Synthetic Parallel Training Data Size in Sentence Pairs: 80000000 Monolingual Training Data in Sentences: 150000000 Processing Tools Used: Tokenizer, Word Aligner (e.g. fast_align or GIZA++), Language detection (e.g. for data cleanup) Toolkit Used: Marian, fairseq(-py) Batch size: 1500 Features of your model development: Data filtering, Data selection, Knowledge distillation, Back-translation with sampling, Iterative back-translation, Forward translation for synthetic data, Oversampling, Ensembling, Averaging, Right-to-left reranking, Fine-tuning for domain adaptation Number of Systems Ensembled/Averaged: 4
HW-TSC	en-de	Multilingual MT System: No. Token Unit Type Used: Moses Tokenizer, spm Vocabulary Size: 32k True Parallel Training Data Size in Sentence Pairs: 79M Synthetic Parallel Training Data Size in Sentence Pairs: 300M Monolingual Training Data in Sentences: en 300M,de 300M Processing Tools Used: Tokenizer, Word Aligner (e.g. fast_align or GIZA++), Language detection (e.g. for data cleanup) Toolkit Used: Marian, fairseq(-py), Moses Features of your model development: Data filtering, Data selection, Knowledge distillation, Back-translation with sampling, Iterative back-translation, Forward translation for synthetic data, Ensembling, Averaging, Fine-tuning for domain adaptation Features of your model structure: Dropout Number of Systems Ensembled/Averaged: 4 ensembled, 3 averaged. Wallclock training time: max_token=500000, max_step=50000
HW-TSC	de-en	Multilingual MT System: No. Token Unit Type Used: Unigram (as in https://github.com/google/sentencepiece), Moses Tokenizer Vocabulary Size: 32K True Parallel Training Data Size in Sentence Pairs: 79M Synthetic Parallel Training Data Size in Sentence Pairs: 300M Monolingual Training Data in Sentences: en 300M, de 300M+ Processing Tools Used: Tokenizer, Word Aligner (e.g. fast_align or GIZA++), Language detection (e.g. for data cleanup) Toolkit Used: Marian, fairseq(-py) Batch size: max_token=500000 Features of your model development: Data filtering, Data selection, Knowledge distillation, Back-translation with sampling, Iterative back-translation, Forward translation for synthetic data, Ensembling, Averaging, Fine-tuning for domain adaptation Features of your model structure: Dropout Number of Systems Ensembled/Averaged: ensembled: 4, average: 3 Wallclock training time: step 50000

C.19 ICL (no associated paper)

No brief description provided.

C.20 IICT-YVERDON

IICT-Yverdon presents the systems submitted by our team from the Institute of ICT (HEIG-VD / HES-SO) to the Unsupervised MT and Very Low Resource Supervised MT task. We first study a baseline system using a Transformer architecture, using the Upper Sorbian (HSB) / German data from the 2020 edition of the task. We quantify the improvements brought by additional techniques such as back-translation of large German corpora and parent-language initialization using Czech-German data, and show that each of these is beneficial, and helps to reach scores that are comparable to more sophisticated systems from the 2020 task. We then present the application of this system to the 2021 task for low-resource supervised HSB-DE translation, in both directions. Finally, we present a contrastive system for HSB-DE in both directions, and for unsupervised German to Lower Sorbian (DSB) translation, which uses multi-task training with various training schedules to improve over the baseline. More specifically, we present a baseline system using a Transformer architecture, which uses back-translation of large German corpora and parent-language initialization using Czech-German data. We submit translations from this system for low-resource supervised HSB-DE, in both directions. We also present a contrastive system that makes use as well of back-translation and Czech-German initialization, and also multi-task training, in which we first train Czech-German systems by giving them different denoising tasks, together with translation, in increasing order of complexity. Afterwards, we first present the child systems with denoising tasks, and later introduce translation. Finally, we train different models with some changes in their training setups that we use for ensembling, in order to maximize diversity among the models.

C.21 IIE-MT (no associated paper)

No brief description provided.

C.22 ILLINI (Le et al., 2021)

Illini team presents an end-to-end NMT pipeline for the Japanese ↔ English news translation task using Transformer models and techniques such as politeness and formality tagging, back-translation, model ensembling, and n-best reranking to improve our translation systems.

C.23 KWAINLP (no associated paper)

No brief description provided.

C.24 LAN-BRIDGE-MT (no associated paper)

No brief description provided.

C.25 LISN (Xu et al., 2021)

LISN's systems for DE↔FR use Transformer-big model with the "priming" based on a prior retrieval step, which looks for similar sentences (in source and target) to prime a similar translation. These techniques aim to perform some unsupervised domain transfer, which is one of the challenge of this task. Our system only uses the data provided for the task (bilingual and backtranslated monolingual data) and are thus constrained submissions. They are built using the fairseq toolkit.

LISN	de-fr, fr-de	Multilingual MT System: No. Basic System Classification: Seq2seq Transformer Style [Vaswani+2017] (self-attention, ...) Token Unit Type Used: BPE (as in https://github.com/rsennrich/subword-nmt), Moses Tokenizer Processing Tools Used: Tokenizer, Language detection (e.g. for data cleanup) Toolkit Used: fairseq(-py) Batch size: 4096

C.26 MACHINE-TRANSLATION (no associated paper)
No brief description provided.

C.27 MANIFOLD (no associated paper)
No brief description provided.

C.28 MIDEIND (Jónsson et al., 2021)

We fine-tuned a sentence-level mBART25 model on the en-is/is-en translation task using a filtered version of the ParIce parallel corpus and a back-translated corpus of roughly 30 million sentence pairs per translation direction. The back-translated corpus was generated via iterative back-translation using a Transformer-base model and a final iteration using the mBART25 translation model. Miðeind is an Icelandic startup company focusing on NLP and AI applications for the Icelandic language.

C.29 MISS (Li et al., 2021b)
No brief description provided.

C.30 MOVELIKEAJAGUAR (no associated paper)
No brief description provided.

C.31 MS-EGDC (Hendy et al., 2021)

We develop NMT for low resource language pairs Bengali to/from Hindi, English to/from Hausa and Xhosa to/from Zulu. We use constrained resources provided by the organizers. The main idea is to train a multi-lingual model with a multi-task objective using both parallel and monolingual data. This model is then used to forward and backward translate monolingual and parallel data (the latter is known as knowledge distillation). The resulting synthetic data is then used to train bilingual MT models for each language pair. The best multi-lingual and multi-task models are then combined with the best bilingual model for each pair using a novel transformer-based method.

C.32 NIUTRANS (Zhou et al., 2021)
No brief description provided.

C.33 NJUSC-TSC (no associated paper)
No brief description provided.

C.34 NUCLEAR-TRANS (no associated paper)
No brief description provided.

C.35 NVIDIA-NEMO (Subramanian et al., 2021)
No brief description provided.

C.36 P3AI (Zhao et al., 2021)
No brief description provided.

C.37 SMU (no associated paper)

No brief description provided.

C.38 TALP-UPC (Escolano et al., 2021)

No brief description provided.

C.39 TRANSSION

This paper describes the submission systems of TRANSSION for WMT21 . We participated in 6 translation directions including Hindi ↔ Bengali, Zulu ↔ Xhosa and English ↔ Hausa in both directions. Our systems are based on Google's Transformer model architecture, into which we integrated the most recent features from the academic research. We also employed most techniques that have been proven effective during the past WMT years, such as Multi-Lingual Training, Back Translation, In-domain Finetuning, Transfer Learning, ensemble and Reranking.

TRANSSION	common	Multilingual MT System: No. Token Unit Type Used: Custom Tokenizer, BPE (as in https://github.com/rsennrich/subword-nmt) Vocabulary Size: 50000 Processing Tools Used: Tokenizer, Shallow Dependency Parser (UD), Shallow Consituency Parser, Word Aligner (e.g. fast_align or GIZA++), Language detection (e.g. for data cleanup) Batch size: 6144 Document-level training: No document-level: Our system processes each segment independently. Number of Systems Ensembled/Averaged: 5 Number of GPUs Used Concurrently: 1
TRANSSION	bn-hi	Basic System Classification: Seq2seq Transformer Style [Vaswani+2017] (self-attention, ...), Hybrid Monolingual Training Data in Sentences: 44,035,924 Monolingual Training Data in Words: 329,604,211,372,512,000 Toolkit Used: Custom in Tensorflow, Custom in Keras (whatever is below it) Features of your model development: Data filtering, Data selection, Back-translation with sampling, Iterative back-translation, Forward translation for synthetic data, Extra languages used beyond those listed above (e.g. some form of pivoting or multi-lingual training), Ensembling, Averaging, Right-to-left reranking, Target-to-source reranking, Fine-tuning for domain adaptation Features of your model structure: Dropout, Tied source and target word embeddings, Residual adapters Pre-trained parts of models: Pre-trained word embeddings Wallclock training time: 12hours Number of contrastive configurations used: 15
TRANSSION	xh-zu, zu-xh, bn-hi, hi-bn, ha-en, en-ha	Basic System Classification: Seq2seq Transformer Style [Vaswani+2017] (self-attention, ...) Toolkit Used: Custom in Tensorflow Features of your model development: Data filtering, Data selection, Back-translation with sampling, Iterative back-translation, Forward translation for synthetic data, Oversampling, Extra languages used beyond those listed above (e.g. some form of pivoting or multi-lingual training), Ensembling, Averaging, Right-to-left reranking, Target-to-source reranking, Fine-tuning for domain adaptation Features of your model structure: Dropout, Tied source and target word embeddings Wallclock training time: 12 hours

C.40 TWB

We developed a bidirectional transformer-based system for Hausa-English news translation task. In our paper we give an overview of the data available including the 15,000 hand-crafted parallel dataset which was created internally. Our best systems achieved 17.1 and 12.3 BLEU on EN-HA and HA-EN directions on the task test sets, respectively.

TWB	common	Multilingual MT System: No. Basic System Classification: Seq2seq Transformer Style [Vaswani+2017] (self-attention, ...) Token Unit Type Used: BPE (as in https://github.com/rsennrich/subword-nmt) Vocabulary Size: 50,000 True Parallel Training Data Size in Sentence Pairs: 806345 True Parallel Training Data Size in Words: 10697192(en), 11405851(ha) Toolkit Used: OpenNMT-py Batch size: 4096 tokens Features of your model development: Data filtering, Data selection, Back-translation with sampling, Ensembling, Averaging, Fine-tuning for domain adaptation Features of your model structure: Dropout Document-level training: No document-level: Our system processes each segment independently. Number of Systems Ensembled/Averaged: Averaged up to 8 models Number of GPUs Used Concurrently: 2 Number of contrastive configurations used: 1
TWB	en-ha	Synthetic Parallel Training Data Size in Sentence Pairs: 567231 Synthetic Parallel Training Data Size in Words: 25495541(ha), 23815542(en) Monolingual Training Data in Sentences: Only the 567231 sentence dataset that were machine translated to make synthetic data Monolingual Training Data in Words: 25495541 Wallclock training time: 24 hours
TWB	ha-en	Synthetic Parallel Training Data Size in Sentence Pairs: 1,000,000 Synthetic Parallel Training Data Size in Words: 11442297(en), 13188160(ha) Monolingual Training Data in Sentences: Only the 1,000,000 sentence dataset that were machine translated to make synthetic data Monolingual Training Data in Words: 11442297 Wallclock training time: 36 to 48 hours

C.41 UEDIN (Chen et al., 2021; Pal et al., 2021)

UEdin's bn-hi and hi-bn systems use models trained on constrained parallel data to back-translate all of the provided monolingual data. New transformer models are then pre-trained on back-translated data, and fine-tuned on parallel data. A second stage of fine-tuning is done on training data that is in-domain, which is extracted in a number of ways, including n-gram matching, TF-IDF similarity, and language model scoring with the validation set. Finally, multiple models fine-tuned in different ways are ensembled to generate the final translations.

UEdin's approach to de↔en started with rule-based and dual conditional cross-entropy filtering of the provided corpora. All models were trained on a mix of parallel and back-translated data, and further trained on parallel sentences only. Specifically for en→de, we trained the model on additional title-cased sentences. The models were then fine-tuned on previous WMT test sets. We ensembled 5 models for en→de and 6 for de→en. During inference, each test instance was split at sentence-level, translated, and then concatenated.

UEDIN	common	Multilingual MT System: No. Basic System Classification: Seq2seq Transformer Style [Vaswani+2017] (self-attention, ...) Token Unit Type Used: Unigram (as in https://github.com/google/sentencepiece) Toolkit Used: Marian Document-level training: No document-level: Our system processes each segment independently. Number of GPUs Used Concurrently: 4

UEDIN	bn-hi	Vocabulary Size: 32000 True Parallel Training Data Size in Sentence Pairs: 2036669 True Parallel Training Data Size in Words: 24797974 Synthetic Parallel Training Data Size in Sentence Pairs: 248828890 Synthetic Parallel Training Data Size in Words: hi (monolingual, target side): 4368794315 bn (back-translated, source side): 3287105444 Monolingual Training Data in Sentences: 248828890 Monolingual Training Data in Words: 4368794315 Processing Tools Used: Tokenizer, Language detection (e.g. for data cleanup) Other Processing Tools Used: Sentence splitter Batch size: Dynamic Features of your model development: Data filtering, Data selection, Ensembling, Fine-tuning for domain adaptation, Back-translation with beam search Number of Systems Ensembled/Averaged: 5 Wallclock training time: 40 (6 * 4 for model ensemble for back-translation + the rest for the final model) Number of contrastive configurations used: 30
UEDIN	hi-bn	Vocabulary Size: 32000 True Parallel Training Data Size in Sentence Pairs: 2036669 True Parallel Training Data Size in Words: 24797974 Synthetic Parallel Training Data Size in Sentence Pairs: 59736357 Synthetic Parallel Training Data Size in Words: bn (monolingual, target side): 873200873 hi (back-translated, source side): 1044281945 Monolingual Training Data in Sentences: 59736357 Monolingual Training Data in Words: 873200873 Processing Tools Used: Tokenizer, Language detection (e.g. for data cleanup) Other Processing Tools Used: Sentence splitter Batch size: Dynamic Features of your model development: Data filtering, Data selection, Forward translation for synthetic data, Ensembling, Fine-tuning for domain adaptation, Back-translation with beam search Number of Systems Ensembled/Averaged: 8 Wallclock training time: 50 (8 * 4 for model ensemble for back-translation + the rest for the final model) Number of contrastive configurations used: 30
UEDIN	de-en	Vocabulary Size: 32k True Parallel Training Data Size in Sentence Pairs: 66530788 Synthetic Parallel Training Data Size in Sentence Pairs: 91033109 Processing Tools Used: Language detection (e.g. for data cleanup) Other Processing Tools Used: fastText for language identification Features of your model development: Data filtering, Back-translation with greedy decoding, Back-translation with sampling, Ensembling, Fine-tuning for domain adaptation Features of your model structure: Dropout, Tied source and target word embeddings Pre-trained parts of models: Did not use Number of Systems Ensembled/Averaged: 6 Number of contrastive configurations used: N/A
UEDIN	en-de	Vocabulary Size: 32k True Parallel Training Data Size in Sentence Pairs: 66530788 Synthetic Parallel Training Data Size in Sentence Pairs: 146216106 Processing Tools Used: Language detection (e.g. for data cleanup) Other Processing Tools Used: fastText for language identification Features of your model development: Data filtering, Back-translation with greedy decoding, Back-translation with sampling, Ensembling, Fine-tuning for domain adaptation Features of your model structure: Dropout, Tied source and target word embeddings Pre-trained parts of models: did not use Number of Systems Ensembled/Averaged: 5 Wallclock training time: 274 hours Number of contrastive configurations used: N/A

C.42 UF (no associated paper)

No brief description provided.

C.43 VOLCTRANS (Qian et al., 2021)

VOLCTRANS-AT VolcTrans-AT's submission is described in the respective paper (Qian et al., 2021).

VOLCTRANS-GLAT VolcTrans-GLAT's submission is a non-autoregressive model equipped with our recent technique of "glancing transformer" (Qian et al., 2020, to appear in ACL 2021).

VOLCTRANS	common	Multilingual MT System: No. True Parallel Training Data Size in Sentence Pairs: 75M Processing Tools Used: Tokenizer, Word Aligner (e.g. fast_align or GIZA++), Language detection (e.g. for data cleanup) Document-level training: No document-level: Our system processes each segment independently.
VOLCTRANS-AT	de-en	Basic System Classification: Seq2seq Transformer Style [Vaswani+2017] (self-attention, ...) Token Unit Type Used: BPE (as in https://github.com/rsennrich/subword-nmt), Moses Tokenizer Vocabulary Size: 12000 Synthetic Parallel Training Data Size in Sentence Pairs: 110M Monolingual Training Data in Sentences: 0 Other Processing Tools Used: n/a Toolkit Used: fairseq(-py), Custom in Pytorch, Custom in Keras (whatever is below it), Moses Batch size: 125k-256k Features of your model development: Data filtering, Data selection, Knowledge distillation, Iterative back-translation, Forward translation for synthetic data, Ensembling, Fine-tuning for domain adaptation Features of your model structure: Dropout, Tied source and target word embeddings Number of Systems Ensembled/Averaged: 9 Number of GPUs Used Concurrently: 16 Wallclock training time: 2 days Other comments: 3
VOLCTRANS-GLAT	de-en	Basic System Classification: Non-Autoregressive Transformer Token Unit Type Used: Unigram (as in https://github.com/google/sentencepiece), Moses Tokenizer Vocabulary Size: 32000 Synthetic Parallel Training Data Size in Sentence Pairs: 100M Monolingual Training Data in Sentences: 0 Toolkit Used: fairseq(-py), Custom in Pytorch, Moses Batch size: 256k Features of your model development: Data filtering, Data selection, Knowledge distillation, Iterative back-translation, Forward translation for synthetic data, Ensembling, Right-to-left reranking Features of your model structure: Dropout Number of Systems Ensembled/Averaged: 3 Number of GPUs Used Concurrently: 32 Wallclock training time: 3 days Number of contrastive configurations used: 6
VOLCTRANS-AT	en-de	Basic System Classification: Seq2seq Transformer Style [Vaswani+2017] (self-attention, ...) Token Unit Type Used: BPE (as in https://github.com/rsennrich/subword-nmt), Moses Tokenizer Vocabulary Size: 12000 Synthetic Parallel Training Data Size in Sentence Pairs: 110M Monolingual Training Data in Words: 0 Toolkit Used: fairseq(-py), Custom in Pytorch, Custom in Keras (whatever is below it), Moses Batch size: 125k-256k Features of your model development: Data filtering, Data selection, Knowledge distillation, Iterative back-translation, Forward translation for synthetic data, Ensembling Features of your model structure: Dropout, Tied source and target word embeddings Number of Systems Ensembled/Averaged: 3 Number of GPUs Used Concurrently: 16 Wallclock training time: 3 days Number of contrastive configurations used: 3
VOLCTRANS-GLAT	en-de	Basic System Classification: Non-Autoregressive Transformer Token Unit Type Used: Unigram (as in https://github.com/google/sentencepiece), Moses Tokenizer Vocabulary Size: 32000 Synthetic Parallel Training Data Size in Sentence Pairs: 100M Monolingual Training Data in Sentences: 0 Toolkit Used: fairseq(-py), Custom in Pytorch Batch size: 256k Features of your model development: Data filtering, Data selection, Knowledge distillation, Iterative back-translation, Fine-tuning for domain adaptation Features of your model structure: Dropout Number of GPUs Used Concurrently: 32 Wallclock training time: 3 days Number of contrastive configurations used: 6

C.44 WATERMELON

We only truly participated de-en direction using constraint settings. For other directions, we submit results from online translators (mainly from DeepL) just in order to see the performance.

WATERMELON	de-en	Multilingual MT System: No. Basic System Classification: Seq2seq Transformer Style [Vaswani+2017] (self-attention, ...) Token Unit Type Used: BPE (as in https://github.com/rsennrich/subword-nmt) Vocabulary Size: 32000 True Parallel Training Data Size in Sentence Pairs: 45M Synthetic Parallel Training Data Size in Sentence Pairs: 65M Processing Tools Used: Tokenizer, Word Aligner (e.g. fast_align or GIZA++), Language detection (e.g. for data cleanup) Other Processing Tools Used: Truecaser Toolkit Used: fairseq(-py) Features of your model development: Data filtering, Data selection, Knowledge distillation, Back-translation with greedy decoding, Back-translation with sampling, Iterative back-translation, Forward translation for synthetic data, Ensembling, Averaging, Right-to-left reranking, Target-to-source reranking, Fine-tuning for domain adaptation Features of your model structure: Dropout, Tied source and target word embeddings Number of Systems Ensembled/Averaged: 15

C.45 WECHAT-AI (Zeng et al., 2021)

We have participated in the WMT 2021 shared news translation task on English-to-Chinese, English-to-Japanese, Japanese-to-English and English-to-German. Our systems are based on the Transformer (Vaswani et al., 2017) with some effective variants, such as mixed-aan model, dual-attention model, weighted-aan model, talking-heads attention model, etc. In our experiments, we employ data selection, several synthetic data generation approaches, advanced finetuning approaches and self-bleu based model ensemble. Our constrained systems achieve 36.9, 46.9, 27.8 and 31.3 case-sensitive BLEU scores on English-to-Chinese, English-to-Japanese, Japanese-to-English and English-to-German, respectively. The BLEU scores of English-to-Chinese, English-to-Japanese and Japanese-to-English are the highest among all submissions, and that of English-to-German ranks the second. Additionally, one of our submissions on English-to-Chinese also achieves the highest chrF score 0.344.

WECHAT-AI	common	Multilingual MT System: No. Basic System Classification: Seq2seq Transformer Style [Vaswani+2017] (self-attention, ...) Token Unit Type Used: BPE (as in https://github.com/rsennrich/subword-nmt) Processing Tools Used: Tokenizer, Word Aligner (e.g. fast_align or GIZA++), Language detection (e.g. for data cleanup) Batch size: 65536 tokens Features of your model structure: Dropout Document-level training: No document-level: Our system processes each segment independently.
WECHAT-AI	en-de	Toolkit Used: fairseq(-py) Features of your model development: Data filtering, Data selection, Knowledge distillation, Back-translation with sampling, Forward translation for synthetic data, Ensembling, Fine-tuning for domain adaptation Number of Systems Ensembled/Averaged: 6
WECHAT-AI	en-ja	Vocabulary Size: en: 34981, ja: 48519 True Parallel Training Data Size in Sentence Pairs: 12339352 True Parallel Training Data Size in Words: en: 310739662, ja: 379286579 Toolkit Used: OpenNMT-py Features of your model development: Data filtering, Data selection, Knowledge distillation, Back-translation with sampling, Iterative back-translation, Ensembling, Fine-tuning for domain adaptation Number of Systems Ensembled/Averaged: 8

WECHAT-AI	ja-en	Vocabulary Size: en: 34981, ja: 48519 True Parallel Training Data Size in Sentence Pairs: 12339352 True Parallel Training Data Size in Words: en: 310739662, ja: 310739662 Toolkit Used: OpenNMT-py Features of your model development: Data filtering, Data selection, Knowledge distillation, Back-translation with sampling, Forward translation for synthetic data, Ensembling, Fine-tuning for domain adaptation Number of Systems Ensembled/Averaged: 15
WECHAT-AI	en-zh	Vocabulary Size: en: 38038, zh: 47038 True Parallel Training Data Size in Sentence Pairs: 31076375 True Parallel Training Data Size in Words: en: 784141085, zh: 749465141 Toolkit Used: fairseq(-py) Features of your model development: Data filtering, Data selection, Knowledge distillation, Back-translation with sampling, Iterative back-translation, Forward translation for synthetic data, Ensembling, Fine-tuning for domain adaptation Number of Systems Ensembled/Averaged: 4

C.46 WINDFALL (no associated paper)

No brief description provided.

C.47 XMU (no associated paper)

No brief description provided.

C.48 YYDS (no associated paper)

No brief description provided.

C.49 ZENGHUIMT (Zeng, 2021)

No brief description provided.

ZENGHUIMT	en-zh, zh-en	Multilingual MT System: No. Basic System Classification: Seq2seq Transformer Style [Vaswani+2017] (self-attention, ...) Token Unit Type Used: Custom Tokenizer, BPE (as in https://github.com/rsennrich/subword-nmt) Vocabulary Size: 45467 True Parallel Training Data Size in Sentence Pairs: 5600583 True Parallel Training Data Size in Words: 88573016 Synthetic Parallel Training Data Size in Sentence Pairs: 23428568 Monolingual Training Data in Sentences: 23428568 Toolkit Used: THUMT Batch size: 15000 Features of your model development: Data filtering, Data selection, Iterative back-translation, Ensembling Features of your model structure: Dropout, Tied source and target word embeddings Document-level training: No document-level: Our system processes each segment independently. Number of Systems Ensembled/Averaged: 4 Number of GPUs Used Concurrently: 1 Wallclock training time: three days

C.50 ZMT (no associated paper)

No brief description provided.