Findings of the 2020 Conference on Machine Translation (WMT20)

Loïc Barrault University of Sheffield
Magdalena Biesialska UPC
Ondřej Bojar Charles University
Marta R. Costa-jussà UPC
Christian Federmann Microsoft Cloud + AI
Yvette Graham Trinity College Dublin
Roman Grundkiewicz Microsoft
Barry Haddow University of Edinburgh
Matthias Huck SAP SE
Eric Joanis NRC
Tom Kocmi Microsoft
Philipp Koehn JHU
Chi-kiu Lo NRC
Nikola Ljubešić Josef Stefan Institute
Christof Monz University of Amsterdam
Makoto Morishita NTT
Masaaki Nagata NTT
Toshiaki Nakazawa University of Tokyo
Santanu Pal WIPRO AI
Matt Post JHU
Marcos Zampieri Rochester Institute of Technology

Abstract

This paper presents the results of the news translation task and the similar language translation task, both organised alongside the Conference on Machine Translation (WMT) 2020. In the news task, participants were asked to build machine translation systems for any of 11 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 task, participants built machine translation systems for translating between closely related pairs of languages.

1 Introduction

The Fifth Conference on Machine Translation (WMT20)[1] was held online with EMNLP 2020 and hosted a number of shared tasks on various aspects of machine translation. This conference built on 14 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, 2018; Barrault et al., 2019).

This year we conducted several official tasks. We report in this paper on the news and similar translation tasks. Additional shared tasks are described in separate papers in these proceedings:

- automatic post-editing (Chatterjee et al., 2020)
- biomedical translation (Bawden et al., 2020b)
- chat translation (Farajian et al., 2020)
- lifelong learning (Barrault et al., 2020)
- metrics (Mathur et al., 2020)
- parallel corpus filtering (Koehn et al., 2020)
- quality estimation (Specia et al., 2020a)
- robustness (Specia et al., 2020b)
- unsupervised and very low-resource translation (Fraser, 2020)

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 22 translation directions this year, with translation between English and each of Chinese, Czech, German and Russian, as well as French to and from German being repeated from last year, and English to and from Inuktitut, Japanese, Polish and Tamil being new for this year. Furthermore, English to and from Khmer and Pashto were included, using the same test sets as in the corpus filtering task. 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. In addition, we used Mechanical Turk to collect further evaluations, as well as a pool of linguists. This year, the official manual evaluation metric is again based on judgments of adequacy on a 100-point scale, a method (known

[1] http://www.statmt.org/wmt20/

Proceedings of the 5th Conference on Machine Translation (WMT), pages 1–55

as "direct assessment") that we explored in the previous years with convincing results in terms of the trade-off between annotation effort and reliable distinctions between systems.

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).

In order to gain further insight into the performance of individual MT systems, we organized a call for dedicated "test suites", each focusing on some particular aspect of translation quality. A brief overview of the test suites is provided in Section 4.

Following the success of the first Similar Language Translation (SLT) task at WMT 2019 and the interest of the community in this topic (Costa-jussà et al., 2018; Popović et al., 2020), we organize a second iteration of the SLT task at WMT 2020. The goal of the shared task is to evaluate the performance of state-of-the-art MT systems on translating between pairs of closely-related languages from the same language family. SLT 2020 features five pairs of similar languages from three language families: Indo-Aryan (Hindi and Marathi), Romance (Catalan, Spanish, and Portuguese), and South-Slavic (Croatian, Serbian, and Slovene). Translations were evaluated in both directions using three automatic metrics: BLEU, RIBES, and TER. Results and main findings of the SLT shared task are discussed in Section 5.

2 News Translation Task

This recurring WMT task assesses the quality of MT on news domain text. As in the previous year, we included Chinese, Czech, German and Russian (into and out of English) as well as French-German. New language pairs for this year were Inuktitut, Japanese, Polish and Tamil (to and from English). We also included the two language pairs from the corpus filtering task (Pashto→English and Khmer→English), to give participants the opportunity to build and test MT systems using the large noisy corpora released for that task.

2.1 Test Data

As in previous years, the test sets consist (as far as possible) 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 test sets differed along several dimensions, which we list in Table 1. The differing aspects of the test sets are as follows:

Domain Most test sets are drawn from the "news" domain, which means the source texts were extracted from online news websites, and the translations were produced specifically for the task. The Pashto→English and Khmer→English test sets were drawn from wikipedia and, as last year, the French↔German test sets concentrated on EU-related news.

Due to limited resources and data available, the Inuktitut↔English test sets contain document- and sentence-aligned data collected from two domains: news and parliamentary. The news data were extracted from the Nunatsiaq News online news website. The parliamentary data were debates from the Nunavut Hansard that are more recent than the training corpus.

Development? For new languages we released a development set, produced in the same way as the test set.

Sentence-split? For some pairs, we did not sentence-split the source texts. In these cases, we extracted the text from the HTML source with paragraph breaks retained, and asked translators to maintain only the paragraph breaks. This was done in order to try to improve the quality of the human translation by allowing the translators more freedom. Some analysis of the paragraph-split pairs is presented in Section 2.1.1.

Directional? For most language pairs the source-side of the test set is the original, and the target-side of the test set is the translation. This is in contrast to the situation up until 2018 when our test sets were constructed from both "source-original" and "target-original" parts. Where a

[2] http://statmt.org/wmt20/results.html

development set is provided, it is a mixture of both "source-original" and "target-original" texts, in order to maximise its size, although the original language is always marked in the sgm file, except for Inuktitut↔English. The consequences of directionality in test sets has been discussed recently in the literature (Freitag et al., 2019; Laubli et al., 2020; Graham et al., 2020), and the conclusion is that it can have an effect on detrimental effect on the accuracy of system evaluation. We use "source-original" parallel sentences wherever possible, on the basis that it is the more realistic scenario for practical MT usage.

Exception: the test sets for the two Inuktitut↔English translation directions contain the same data, without regard to original direction. For most news text in the test and development sets, English was the original language and Inuktitut the translation, while the parliamentary data mixes the two directions.

The origins of the news test documents is shown in Table 5, and the size of the test sets in terms of sentence pairs and words is given in Figure 4. We generally aimed for 1000 sentences for a new language pair, and 2000 sentences for a previously used language pair (since there was no need to create a development set for a previously-used language pair). For test sets where the source was not sentence-split (see below) we aimed for an equivalent to 2000 sentences, but in running words.

In order to improve the consistency and quality of the test set translations, this year we prepared common translator briefs to be sent to each agency we used. We show the translator briefs in Appendix B (for sentence-split sources) and Appendix C (for paragraph-split sources).

2.1.1 Paragraph-split Test Sets

For the language pairs English↔Czech, English↔German and English→Chinese, we provided the translators with paragraph-split texts, instead of sentence-split texts. We did this in order to provide the translators with greater freedom and, hopefully, to improve the quality of the translation. Allowing translators to merge and split sentences removes one of the "translation shifts" identified by Popovic (2019), which can make translations create solely for MT evaluation different from translations produced for other purposes.

We first show some descriptive statistics of the source texts, for Czech, English and German, in Table 2, where we used the Moses sentence splitter (Koehn et al., 2007) to provide sentence boundaries. We can see that the number of sentences per paragraph is much lower for English, where in fact 70% of paragraphs only have single sentence. For Czech and German, the mean sentences per paragraph is quite similar (2.62 vs. 2.52).

The main question though, is whether translators tended to preserve the sentence structure when translating. To determine this, we split both source paragraphs and translations into sentence, and aligned them using hunalign (Varga et al., 2005) with the bitextor dictionaries (Esplà-Gomis, 2009). In Table 4 we show the counts of 1-1 sentence alignments, as well as cases where the translator merged or split neighbouring sentences. Note that these counts are approximate, since they are affected by errors in the automatic splitting and alignment.

Looking through examples of merges and splits, we see that most of them are relatively simple changes, where the translator has merged to clauses into a sentence, or split a sentence to clauses. Examples of such merges and splits are shown in Table 3, where the first and second are simple merges or splits, whereas the third is a rare case of more complex reordering. We leave a detailed analysis of the translators' treatment of paragraph-split data for future work.

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. These included Wikimatrix (Schwenk et al., 2019), which was added for all language pairs where it was available. The news commentary and europarl corpora that we have been using since the earliest news task now have "data sheets", describing the data sets in standardised format (Costa-jussà et al., 2020).

For Tamil-English, we additionally included some recently crawled multilingual parallel corpora from Indian government websites (Haddow and Kirefu, 2020; Siripragada et al., 2020), the Tanzil corpus (Tiedemann, 2009), the Pavlick dic-

Europarl Parallel Corpus

	Czech ↔ English		German ↔ English		Polish↔ English		German ↔ French	
Sentences	645,241		1,825,745		632,435		1,801,076	
Words	14,948,900	17,380,340	48,125,573	50,506,059	14,691,199	16,995,232	47,517,102	55,366,136
Distinct words	172,452	63,289	371,748	113,960	170,271	62,694	368,585	134,762

News Commentary Parallel Corpus

	Czech ↔ English		German ↔ English		Russian ↔ English	
Sentences	248,927		361,735		308,853	
Words	5,570,734	6,156,063	9,199,170	9,127,331	7,867,940	8,200,081
Distinct words	174,952	70,115	206,506	83,701	201,616	80,219

	Chinese ↔ English		Japanese ↔ English		German ↔ French	
Sentences	312,489		1,818		276,637	
Words	–	7,939,817	–	44,418	7,148,178	8,703,088
Distinct words	–	76,013	–	6,165	178,453	85,189

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		Polish ↔ English	
Sentences	34,371,306		5,345,693		6,577,804	
Words	767,321,987	813,326,217	115,294,152	124,695,776	151,873,495	167,023,296
Distinct Words	8,187,923	4,151,916	1,503,435	1,030,918	1,926,833	1,386,287

	Japanese ↔ English		Russian ↔ English		French ↔ German	
Sentences	10,120,013		12,061,155		7,222,574	
Words	–	274,368,443	182,325,667	210,770,840	145,190,707	123,205,701
Distinct Words	–	2,051,246	2,958,831	2,385,076	1,534,068	2,368,682

	Khmer ↔ English		Pashto ↔ English	
Sentences	4,169,574		1,022,883	
Words	–	77,927,333	14,442,909	13,890,077
Distinct Words	–	1,002,134	365,781	349,261

EU Press Release Parallel Corpus

	Czech ↔ English		German ↔ English		Polish↔ English	
Sentences	452,411		1,631,639		277,984	
Words	7,214,324	7,748,940	26,321,432	27,018,196	6,415,074	6,904,358
Distinct words	141,077	83,733	402,533	197,030	121,451	62,672

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

	Czech ↔ English		German ↔ English		Inuktitut ↔ English		Japanese↔ English	
Sentences	382,336		1,382,681		829		706,012	
Words	916,397	984,247	2,999,545	3,504,013	1213	1213	–	1,867,218
Distinct	206,935	176,156	645,224	547,930	962	938	–	268,391

	Polish ↔ English		Pashto ↔ English		Russian ↔ English	
Sentences	1,006,263		9,869		1,108,789	
Words	2,236,756	2,579,249	20,674	19,519	3,010,302	3,027,765
Distinct	507,571	475,255	9,692	8,899	507,251	434,244

	Tamil↔ English		Chinese↔ English		German↔ French	
Sentences	102,143		836,682		942,017	
Words	237,962	234,380	–	2,267,336	1,989,965	2,363,308
Distinct	72,577	61,267	–	357,440	479,000	423,406

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) for Gujarati.

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

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

Extra Tamil-English Parallel Data

	PIB		MKB		NLPC	
Sentences	60,836		5,744		8,900	
Words	981,352	1,245,455	91,556	114,415	62,041	75,326
Distinct	96,911	35,954	20,697	9,501	13,794	7,087

	UFAL		Tanzil		PMIndia	
Sentences	169,871		93,540		39,526	
Words	3,335,382	4,537,910	2,595,930	2,822,291	604,814	798,406
Distinct	347,874	70,627	27,711	20,282	70,845	25,074

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

Nunavut Hansard Parallel Corpus

	Inuktitut ↔ English	
Sentences	1,301,736	
Words	10,875,086	20,781,805
Distinct	1,594,280	57,691

Opus Corpus

	Khmer ↔ English		Pashto ↔ English	
Sentences	290,049		123,198	
Words	–	4,537,258	889,520	814,064
Distinct	–	52,496	30,583	20,795

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		Polish ↔ English	
Sentences	2,094,650		6,227,188		3,895,992		3,085,946	
Words	34,801,119	39,197,172	113,445,806	118,077,685	–	72,320,248	50,061,388	55,736,716
Distinct	1,068,844	798,095	2,855,263	1,827,785		1,106,529	1,312,825	1,096,411

	Russian ↔ English		Tamil ↔ English		Chinese ↔ English		German ↔ French	
Sentences	5,203,872		240,357		2,595,119		3,350,816	
Words	93,828,313	102,937,537	3,057,383	3,766,628	–	58,615,891	68,249,384	59,422,699
Distinct	2,233,043	1,592,190	392,613	262,094	–	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) for Tamil.

Monolingual Wikipedia Data

	English	Khmer	Pashto	Tamil
Sentences	67,796,935	132,666	76,557	1,669,257
Words	2,277,495,444	–	3,985,596	22,251,345
Distinct words	8,570,978	–	229,040	1,542,047

News Language Model Data

	English	German	Czech	Russian	Japanese
Sentences	233,501,354	333,313,278	81,708,712	93,827,187	3,446,416
Words	5,578,072,595	6,492,440,544	1,429,535,453	1,702,976,902	–
Distinct words	7,590,931	37,274,673	4,890,810	5,199,379	–

	Polish	Chinese	French	Tamil
Sentences	3,788,276	4,724,008	87,063,385	708,500
Words	66,323,590	–	2,105,883,073	9,421,383
Distinct words	725,050	–	3,736,705	536,423

Document-Split News LM Data (not dedudped)

	Czech	English	German
Sentences	114,101,660	486,139,068	654,097,256
Words	1,798,383,105	10,459,366,947	11,097,364,402
Distinct words	4,765,875	7,857,783	24,538,295

Common Crawl Language Model Data

	English	German	Czech	Russian	Polish
Sent.	3,074,921,453	2,872,785,485	333,498,145	1,168,529,851	1,422,729,881
Words	65,104,585,881	65,147,123,742	6,702,445,552	23,332,529,629	40,639,985,955
Dist.	342,149,665	338,410,238	48,788,665	90,497,177	213,298,869

	Chinese	Inuktitut	Tamil	Pashto	French
Sent.	1,672,324,647	296,730	28,828,239	6,558,180	4,898,012,445
Words	–	1,480,611	632,363,004	218,412,919	126,364,574,036
Dist.	–	448,513	16,780,006	23,531,044	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) for Tamil.

Test Sets

	Czech → EN		EN → Czech		German → EN			EN → German		
Lines.	664		1418		785			1418		
Words	30069	39570	50330	47553	35475	38559	38322	50330	53243	53837
Distinct words	10043	6303	7893	12667	7923	5936	5954	7893	10563	10536

	Inuktitut ↔ EN		Tamil → EN		EN → Tamil		Japanese → EN		EN → Japanese		Khmer ↔ EN	
Lines.	2971		997		1000		993		1000		2320	
Words	36710	68111	15402	19716	25176	19749	–	28446	25176	–	–	45220
Distinct words	14531	5719	6183	3519	4971	8139	–	5195	4971	–	–	5315

	Pashto ↔ EN		Polish → EN		EN → Polish		EN → Russian		German ↔ French	
Lines.	2719		1001		1000		2002		1619	
Words	59245	53754	18472	21852	25176	24346	49862	47909	30422	40180
Distinct words	9071	6305	6685	4274	4971	7997	7772	13042	5428	4727

	Chinese → EN			EN → Chinese			Russia → EN		
Lines.	2000			1418			991		
Words	–	74835	74700	50330	–	–	17249	20346	20704
Distinct words	–	8137	8209	7893	–	–	6328	4091	4066

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) for Tamil

Pair	Domain	Development?	Sentence-split?	Directional?	Documents?
Chinese↔English	news	✗	Only zh→en	✓	✓
Czech↔English	news	✗	✗	✓	✓
French↔German	EU related news	✗	✓	✗	✗
German↔English	news	✗	✗	✓	✓
Inuktitut↔English	news and parliament	✓	✓	✗	Only news
Japanese↔English	news	✓	✓	✓	✓
Khmer↔English	wikipedia	✓	✓	✗	✗
Pashto↔English	wikipedia	✓	✓	✗	✗
Polish↔English	news	✓	✓	✓	✓
Russian↔English	news	✗	✓	✓	✓
Tamil↔English	news	✓	✓	✓	✓

Table 1: The characteristics of the test sets for the news tasks. We show the domain that the test set was drawn from, whether or not we released a development set this year, whether the texts were sentence-split before translation, and whether the direction of translation was preserved. For "directional" test sets, the entire source side of the test set was originally written in the source language, and then translated to the target language. Non-directional test sets are a mixture of "source-original" and "target-original" texts. Finally, we record whether or not the test set contained the original document boundaries.

Language	Documents	Paragraphs	Sentences	Words	Sentences per Paragraph
Czech	102	659	1725	25874	2.62
English	130	1418	2043	44018	1.44
German	118	777	1958	31030	2.52

Table 2: Descriptive statistics of paragraph-split source texts. To count the sentences, we applied the Moses sentence splitter to the texts.

Als Rückzieher sei das aber nicht zu verstehen: "Ganz Ägypten ist der Tahrirplatz".	But that should not be understood as a withdrawal. "All of Egypt is Tahrir square."
"Ich fühle mich unglaublich geehrt und demütig, neben JLo die Latino-Community zu repräsentieren. Denn diese hat eine unglaubliche Stärke in den USA", teilte Shakira in einem Video mit.	"I feel incredibly honored and humbled to be next to J. Lo, representing the Latino community that is such an important force in the United States," Shakira shared in a video.
Man könne die Unternehmen zwar nicht von der Umsatzsteuer auf Sachspenden befreien, erklärte das Ministerium auf eine Frage der Grünen-Bundestagsfraktion, über die die Zeitungen der Funke-Mediengruppe am Freitag berichteten. Die Händler könnten aber den Marktwert der unverkäuflichen Retouren so niedrig ansetzen, dass sie keine oder nur wenig Umsatzsteuer zahlen müssten.	Although it is impossible to exempt companies from VAT on donations in kind, retailers could set the market value of unsaleable returns so low that they would need to pay no or only very little VAT, the Ministry explained in response to a question from the Greens parliamentary group, as reported in newspapers of the Funke media group on Friday.

Table 3: Examples of translations where the translator has split or merged the sentences. The third example is one of the rare examples of a non-trivial merging of the sentence (i.e. there is merging accompanied by reordering)

tionaries (Pavlick et al., 2014), a corpus[3] produced by the University of Moratuwa, the HindEnCorp

[3] https://github.com/nlpcuom/English-Tamil-Parallel-Corpus

Pair	Translator	1-1	%age	Merges	Splits	n-m
Czech→English	A	1573	91.2	26	98	1
English→Czech	A	2013	98.5	10	8	1
German→English	A	1913	97.7	13	19	0
	B	1844	94.2	35	43	0
English→German	A	2017	98.7	9	8	0
	B	1816	88.9	12	203	0

Table 4: How the translators treated sentences when translating the paragraph-split texts. We sentence-split and automatically aligned source and translation. We show the number and percentage of sentences which were translated 1-1, as well as the number of times translators merged or split sentences when translating.

(Kunchukuttan et al., 2018) and English and Tamil wikipedia dumps.

The training corpus for Inuktitut↔English is the recently released Nunavut Hansard Inuktitut–English Parallel Corpus 3.0 (Joanis et al., 2020).

For the Japanese↔English tasks, we added several freely available parallel corpora to the training data. It includes JParaCrawl v2.0 (Morishita et al., 2020), a large web-based parallel corpus, Japanese-English Subtitle Corpus (JESC) (Pryzant et al., 2017), the Kyoto Free Translation Task (KFTT) corpus (Neubig, 2011), constructed from the Kyoto-related Wikipedia articles, and TED Talks (Cettolo et al., 2012).

The monolingual data we provided was similar to last year's, with a 2019 news crawl added to all the news corpora. 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.

Training, development, and test data for Pashto↔English and Khmer↔English are shared with the Parallel Corpus Filtering Shared Task (Koehn et al., 2020). The training data mostly comes from OPUS (software localization, Tatoeba, Global Voices), the Bible, and special-prepared corpora from TED Talks and the Jehova Witness web site (JW300). The development and test sets were created as part of the Flores initiative (Guzmán et al., 2019) by professional translation of Wikipedia content with careful vetting of the translations. Please refer the to the Parallel Corpus Filtering Shared Task overview paper for details on these corpora.

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

2.3 Submitted Systems

In 2020, we received a total of 153 submissions. The participating institutions are listed in Table 6 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 4 services), which we anonymized as ONLINE-A,B,G,Z.

This year we introduced a new submission tool, OCELoT[4], replacing the matrix that has been used in most previous editions. Using OCELoT gave us more control over the submission and scoring process, for example we were able to limit the number of test submissions by each team, and we also displayed the submissions anonymously to avoid publishing any automatic scores. A screenshot of OCELoT is shown in Figure 5.

For presentation of the results, systems are treated as either *constrained* or *unconstrained*, depending on whether their models were trained only on the provided data. Since we do not know how they were built, the online systems are treated as unconstrained during the automatic and human evaluations.

In the rest of this section, we provide brief details of the submitted systems, for those where the authors provided such details.

2.3.1 AFRL (Gwinnup and Anderson, 2020)

AFRL-SYSCOMB20 is a system combination consisting of two Marian transformer ensembles, one OpenNMT transformer system and a Moses phrase-based system.

AFRL-FINETUNE is an OpenNMT transformer system fine-tuned on newstest2014-2017.

2.3.2 ARIEL XV (Xv, 2020)

ARIEL XV is a Transformer base model trained with the Sockeye sequence modeling toolkit us-

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

English I	ABC News (2), All Africa (5), Brisbane Times (1), CBS LA (1), CBS News (1), CNBC (3), CNN (2), Daily Express (1), Daily Mail (2), Fox News (1), Gateway (1), Guardian (3), Huffington Post (2), London Evening Standard (2), Metro (2), NDTV (7), RTE (7), Reuters (4), STV (2), Seattle Times (3), The Independent (1), The Local (1), The Scotsman (2), The Sun (1), The Telegraph (1), VOA Zimbabwe (1), news.com.au (4),
English II	ABC News (2), Al Jazeera (1), All Africa (6), Brisbane Times (1), CBS LA (1), CNBC (3), CNN (1), Chicago Defender (1), Daily Express (2), Daily Mail (2), Egypt Independent (1), Euronews (1), Guardian (2), Herald Scotland (1), Huffington Post (6), Kazakh TV (1), LA Times (1), London Evening Standard (3), Metro (1), NDTV (6), One India (2), RTE (1), Reuters (1), Russia Today (1), Seattle Times (1), Sky (1), The Independent (1), The Scotsman (4), The Sun (2), UPI (1), news.com.au (3),
English III	ABC News (5), Al Jazeera (3), All Africa (2), Brisbane Times (2), CBS LA (2), CBS News (3), CNBC (5), CNN (6), Chicago Defender (1), Daily Express (2), Daily Mail (2), Euronews (3), Fox News (5), Gateway (1), Guardian (5), Herald Scotland (1), Huffington Post (8), LA Times (2), London Evening Standard (5), Medical Daily (1), Metro (3), NDTV (7), New Republic (1), New York Times (2), Novinite (3), RTE (3), Reuters (8), Russia Today (7), STV (1), Sciencedaily (2), Seattle Times (12), Sky (3), The Independent (2), The Scotsman (1), The Sun (1), The Telegraph (4), UPI (6),
Chinese	China News (64), Chubun (3), Hunan Ribao (5), International Times (10), Jingji Guancha Bao (1), Macao Government (5), Nhan Dan (9), Nikkei (2), Reuters (2), The Australian (2), UN news (2), Xinhua (46), qq.com (1), tsrus.cn (3),
Czech	Aktualne (6), Blesk (13), Denik (7), E15 (3), Hospodářské Noviny (7), Idnes (10), Lidovky (8), Mediafax (3), Neviditelný Pes (2), Novinky (14), Reflex (1), Respekt (5), Týden (9), Česká Pozice (7), České Noviny (7),
German	Allgemeine Zeitung (2), Braunschweiger Zzeitung (3), Dülmener Zeitung (1), Das Bild (2), Der Spiegel (2), Der Standart (2), Deutsche Welle (2), Die Zeit (3), Echo Online (1), Epoch Times (3), Euronews (2), Frankfurter Allgemeine Zeitung (1), Freie Presse (1), Freitag (1), Giessener Anzeiger (1), Goslarsche Zeitung (1), Handelsblatt (2), Heute (2), In Südthüringen (1), Infranken (1), Junge Freiheit (1), Kurier (4), Lübecker Nachrichten (1), Leipziger Volkszeitung (1), Lippische Landes-Zeitung (2), Mittelbayerische Zeitung (2), Mitteldeutsche Zeitung (3), NTV (6), NZZ (5), Nachrichten (2), Neue Osnabrücker Zeitung (1), Neue Presse (1), Neues Deutschland (1), Norddeutsche Neueste Nachrichten (3), OE24 (1), Onetz (1), Passauer Neue Presse (2), Peiner Allgemeine Zeitung (3), Presse Portal (1), Rhein Zeitung (3), Söster Anzeiger (1), Süddeutsche Zeitung (3), Salzburger Nachrichten (4), Schaumburger Nachrichten (1), Schleswig-Holsteinischer Zeitungsverlag (4), Segeberger Zeitung (3), Solinger Tageblatt (1), Stuttgarter Zeitung (1), Tagesspiegel (3), Tiroler Tageszeitung (7), Vaterland (1), Volksblatt (1), Welt (2), Westfälische Nachrichten (1), Westfälischer Anzeiger (1), Wiesbadener Kurier (1), Yahoo (5),
Inuktitut	Nunatsiaq News (36), Nunavut Hansard (1),
Japanese	Fukui Shimbun (6), Hokkaido Shimbun (6), Ise Shimbun (1), Iwaki Minpo (2), Saga Shimbun (3), Sanyo Shimbun (4), Shizuoka Shimbun (15), Ube nippo Shimbun (1), Yahoo (40), Yamagata Shimbun (2),
Polish	Bankier (5), Gazeta Powiatowa (1), Gazeta Prawna (3), Interia (24), Polityka (1), Rzeczpospolita (4), Super Nowosci (3), Sztafeta (1), TVN24 (7), Tygodnik Zamojski (2), WPROST (7), Wyborcza (1), Zycie Podkarpackie (3),
Russian	Argumenti Nedely (6), Argumenty i Fakty (9), BBC Russian (2), Delovoj Peterburg (2), ERR (2), Ekonomika i Zhizn (1), Fakty i Komentarii (3), Gazeta (4), Interfax (3), Izvestiya (7), Kommersant (4), Komsomolskaya Pravda (4), Lenta (7), Moskovskij Komsomolets (3), Nasha Versiya (1), Novye Izvestiya (1), Parlamentskaya Gazeta (5), Rosbalt (5), Rossiskaya Gazeta (1), Russia Today (3), Russkaya Planeta (1), Sport Express (6), Tyumenskaya Oblast Segodnya (1), Vedomosti (2), Vesti (6), Xinhua (2),
Tamil	Aranda Vikatan (11), Dinamalar (2), Makkal Kural (21), One India (21), Viduthalai (15), news.lk (12),

Table 5: Composition of the test sets. English I was used for English to Japanese, Polish, Russian and Tamil, English II was used additionally for English to Russian, and English III (which was not sentence-split) was translated to Czech, German and Chinese. The same document pairs were used in both directions for Inuktitut↔English. For more details see the XML test files. The `docid` tag gives the source and the date for each document in the test set, and the `origlang` tag indicates the original source language.

Team	Institution
AFRL	Air Force Research Laboratory (Gwinnup and Anderson, 2020)
ARIEL XV	Independent submission (Xv, 2020)
CUNI	Charles University (Popel, 2020, 2018; Kocmi, 2020)
DCU	Dublin City University (Parthasarathy et al., 2020)
DEEPMIND	DeepMind (Yu et al., 2020)
DIDI-NLP	DiDi AI Labs (Chen et al., 2020b)
DONG-NMT	(no associated paper)
ENMT	Indepdendent Submission (Kim et al., 2020)
ETRANSLATION	eTranslation (Oravecz et al., 2020)
FACEBOOK AI	Facebook AI (Chen et al., 2020a)
GRONINGEN	University of Groningen (Roest et al., 2020; Dhar et al., 2020)
GTCOM	Global Tone Communication (Bei et al., 2020)
HELSINKINLP	University of Helsinki and Aalto University (Scherrer et al., 2020a)
HUAWEI TSC	Huawei TSC (Wei et al., 2020a)
IIE	Institute of Information Engineering, Chinese Academy of Sciences (Wei et al., 2020b)
MICROSOFT STC INDIA	Microsoft STC India (Goyal et al., 2020)
NICT-KYOTO	NICT-Kyoto (Marie et al., 2020)
NICT-RUI	NICT-Rui (Li et al., 2020)
NIUTRANS	NiuTrans (Zhang et al., 2020)
NRC	National Research Council Canada (Knowles et al., 2020)
OPPO	OPPO (Shi et al., 2020)
PROMT	PROMT (Molchanov, 2020)
SJTU-NICT	SJTU-NICT (Li et al., 2020)
SRPOL	Samsung Research Poland (Krubiński et al., 2020)
TALP UPC	TALP UPC (Escolano et al., 2020)
TENCENT TRANSLATION	Tencent Translation (Wu et al., 2020b)
THUNLP	NLP Lab at Tsinghua University (no associated paper)
TILDE	Tilde (Krišlauks and Pinnis, 2020)
TOHOKU-AIP-NTT	Tohoku-AIP-NTT (Kiyono et al., 2020)
UBIQUS	Ubiqus (Hernandez and Nguyen, 2020)
UEDIN	University of Edinburgh (Bawden et al., 2020a; Germann, 2020)
UEDIN-CUNI	University of Edinburgh and Charles University (Germann et al., 2020)
UQAM_TANLE	Université du Québec à Montréal (no associated paper)
VOLCTRANS	ByteDance AI Lab (Wu et al., 2020a)
WECHAT	WeChat (Meng et al., 2020)
WMTBIOMEDBASELINE	Baseline System from Biomedical Task (Bawden et al., 2020b)
YOLO	American University of Beirut (no associated paper)
ZLABS-NLP	Zoho Corporation (no associated paper)

Table 6: Participants in the shared translation task. Not all teams participated in all language pairs. 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.

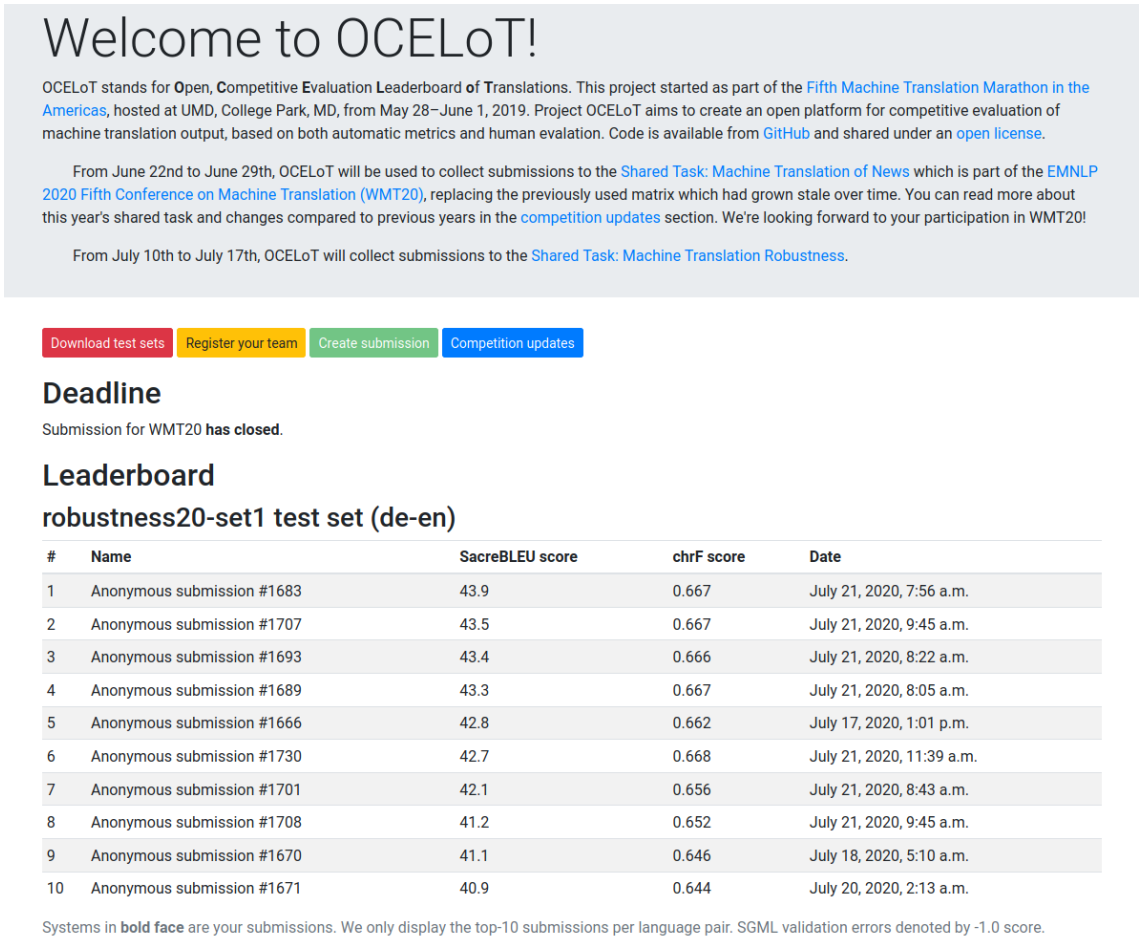

Figure 5: The OCELoT leaderboard tool

ing only the constrained data. The authors experiment with bi-text data filtering, back-translation, rule-based reranking based on translation and language model scores, ensembling several training runs and fine-tuning for sentences similar to the desired domain based on the source side of the test set.

2.3.3 Charles University (CUNI)

CUNI-DocTransformer **(Popel, 2020)** is similar to the sentence-level version (CUNI-T2T-2018, CUBBITT), but trained on sequences with multiple sentences of up to 3000 characters.

CUNI-T2T-2018 **(Popel, 2018),** 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 novel concat-regime backtranslation with checkpoint averaging (Popel et al., 2020), tuned separately for CZ-domain and non CZ-domain articles, possibly handling also translation-direction ("translationese") issues. For cs→en also a coreference preprocessing was used adding the female-gender pronoun where it was pro-dropped in Czech, referring to a human and could not be inferred from a given sentence.

CUNI-Transfer **(Kocmi, 2020)** combines transfer learning from a high-resource language pair Czech–English into the low-resource Inuktitut-English with an additional backtranslation step. Surprising behaviour is noticed when using synthetic data, which can be possibly attributed to a narrow domain of training and test data. The system is the Transformer model in a constrained submission.

CUNI-Transformer **(Popel, 2020)** is similar to the WMT2018 version of CUBBITT, but with 12 encoder layers instead of 6 and trained on CzEng 2.0 instead of CzEng 1.7. The English-Polish version was trained similarly on the provided constrained data.

2.3.4 DCU (Parthasarathy et al., 2020)

DCU participated in the Tamil↔English translations with the Transformer model. Various strate-

gies were tested in order to improve over the baseline, e.g. several techniques of data augmentation and mining as well as a hyperparameter search for better performance of the Transformer model in low-resource scenarios.

2.3.5 DEEPMIND (Yu et al., 2020)

DEEPMIND is a document-level translation system built upon noisy channel factorization. The system optimizes the selection of translations of individual sentences in the document in iterative beam search, replacing sentences with alternative translations. Candidate translations are constructed and later scored using a number of independent components, mainly sequence-to-sequence models trained on large data and highly optimized with techniques of back-translation, distillation, and fine-tuning with in-domain data. MonteCarlo Tree Search decoding and uncertainty estimation are used to improve the robustness of the search for the best sentence translation selection and specialized length models and sentence segmentation help to avoid too short output.

2.3.6 DIDI-NLP (Chen et al., 2020b)

DIDI-NLP is a Transformer model improved with several techniques for model enhancement, including data filtering, data selection, large-scale back-translation, knowledge distillation, fine-tuning, model ensembling, and re-ranking.

Ensembled models include Transformers with relative position attention, larger inner feed-forward network size or reversed source. Multiple domain models based on unsupervised BERT-CLS clusters are used in a dynamically-weighted selection of the next word. The final n-best lists are reranked with MIRA.

2.3.7 DONG-NMT (no associated paper)

No description provided.

2.3.8 ENMT (Kim et al., 2020)

Kim et al. (2020) base their approach on transferring knowledge of domain and linguistic characteristics by pre-training the encoder-decoder model with large amount of in-domain monolingual data through unsupervised and supervised prediction task. The model is then fine-tuned with parallel data and in-domain synthetic data, generated with iterative back-translation. For additional gain, final results are generated with an ensemble model and re-ranked with averaged models and language models.

2.3.9 ETRANSLATION (Oravecz et al., 2020)

ETRANSLATION mainly use the standard training pipeline of Transformer in Marian, using tagged back-translation and other features. Subword units are identified by SentencePiece.

The paper describes the group's concern about computing resources and the practical utility of expensive features like ensembling 2 to 4 bigger models. Techniques that were ineffective in ETRANSLATION's case (e.g. right-to-left model for rescoring English→German or Unicode preprocessing for Japanese→English) are also described.

2.3.10 FACEBOOK AI (Chen et al., 2020a)

FACEBOOK AI focus on low-resource language pairs involving Inuktitut and Tamil using two strategies: (1) exploiting all available data (parallel and monolingual from all languages) and (2) adapting the model to the test domain.

For (1), FACEBOOK AI opt for non-constrained submission, using data derived from CommonCrawl to get strong translation models via iterative backtranslation and self-training and strong language models for noisy channel reranking. Multilingual language models are created using mBART across all the 13 languages of WMT20. For (2), the datasets are tagged for domain, fine-tuned on and further extended with in-domain data.

2.3.11 GRONINGEN

GRONINGEN-ENIU (Roest et al., 2020) investigate the (1) importance of correct morphological segmentation of the polysynthetic Inuktitut, testing rule-based, supervised, semi-supervised as well as unsupervised word segmentation methods, (2) whether or not adding data from a related language (Greenlandic) helps, and (3) whether contextual word embeddings (XLM) improve translation.

GRONINGEN-ENIU use Transformer implemented in Marian with the default setting, improving the performance also with tagged backtranslation, domain-specific data, ensembling and fine-tuning.

GRONINGEN-ENTAM (Dhar et al., 2020) study the effects of various techniques such as linguistically motivated segmentation, back-translation, fine-tuning and word dropout on the English→Tamil News Translation task. Linguis-

tically motivated subword segmentation does not consistently outperform the widely used SentencePiece segmentation despite the agglutinative nature of Tamil morphology. The authors also found that fully-fledged back-translation remains more competitive than its cheaper alternative.

2.3.12 GTCOM (Bei et al., 2020)

GTCOM are unconstrained systems using mBART (Multilingual Bidirectional and Auto-Regressive Transformers), back-translation and forward-translation. Further gains are achieved using rules, language model and RoBERTa model to filter monolingual, parallel sentences and synthetic sentences. The vocabularies are created from both monolingual and parallel data.

2.3.13 HELSINKINLP (Scherrer et al., 2020a)

HELSINKINLP for the Inuktitut-English news translation task focuses on the efficient use of monolingual and related bilingual corpora with multi-task learning as well as an optimized subword segmentation with sampling.

2.3.14 HUAWEI TSC (Wei et al., 2020a)

HUAWEI TSC use Transformer-big with a further increased model size, focussing on standard techniques of careful pre-processing and filtering, back-translation and forward translation, including self-training, i.e. translating one of the sides of the original parallel data. Ensembling of individual training runs is used in the forward as well as backward translation, and single models are created from the ensembles using knowledge distillation. The submission uses THUNMT (Zhang et al., 2017) open-source engine.

2.3.15 IIE (Wei et al., 2020b)

IIE German↔French news translation system is based on the Transformer architecture with some effective improvements. Multiscale collaborative deep architecture, data selection, back translation, knowledge distillation, domain adaptation, model ensemble and re-ranking are employed and proven effective in the experiments.

2.3.16 MICROSOFT STC INDIA (Goyal et al., 2020)

Focusing on English↔Tamil, MICROSOFT STC INDIA experiment with "contact relatedness" of languages, i.e. using Hindi-English data in joint training. Hindi texts first have to be mapped from the Devanagari script to Tamil characters in a lossy but deterministic way. Further gains are obtained from tagged back-translation and other variants of back-translation are also examined (noisification or back-translating with right-to-left models).

Transformer implemented in fairseq is used, with smaller than "base" models due to limited training data.

2.3.17 NICT-KYOTO (Marie et al., 2020)

NICT-KYOTO is a combination of neural machine translation systems processed through n-best list reranking. The systems combined are Transformer-based trained with Marian and Fairseq with and without using tagged back-translation. All the systems are constrained, and the final primary submission is selected on the basis of the BLEU score obtained on the official validation data.

2.3.18 NICT-RUI (Li et al., 2020)

NICT-RUI (Li et al., 2020) NICT-RUI is closely related to SJTU-NICT using large XLM model to improve NMT but the exact relation is unclear.

2.3.19 NIUTRANS (Zhang et al., 2020)

NIUTRANS gain their performance from focussed attention to six areas: (1) careful data preprocessing and filtering, (2) iterative back-translation to generate additional training data, (3) using different model architectures, such as wider and/or deeper models, relative position representation and relative length, to enhance the diversity of translations, (4) iterative knowledge distillation by in-domain monolingual data, (5) iterative fine-tuning for domain adaptation using small training batches, (6) rule-based post-processing of numbers, names and punctuation.

For low-resource language pairs, multi-lingual seed models are used.

2.3.20 NRC (Knowles et al., 2020)

The NRC systems are hybrids of Transformer models trained with Sockeye, with one ensembled system for news domain translation and one for Hansard domain translation. Data was pre-processed with language-specific punctuation and character preprocessing, tokenization, and BPE. They were trained with domain tagging, domain-specific finetuning, ensembles of 3 systems per domain, BPE-dropout (EN-IU), and tagged back-translation (IU-EN).

2.3.21 OPPO (Shi et al., 2020)

OPPO train Marian for some language pairs and fairseq for others, relying on a number of mature techniques including careful corpus filtering, iterative forward and backward translation, fine-tuning on the original parallel data, ensembling of several different models, and complex reranking which uses forward (source-to-target) scorers, backward scorers (target-to-source) and language models (monolingual), each group again building upon ensembles and being applied left-to-right as well as right-to-left.

Each language pair received targeted attention, discussing training data properties, varying the process as needed and choosing from several possible final models.

2.3.22 PROMT (Molchanov, 2020)

PROMT BASELINE TRANSFORMER uses MarianNMT, shared vocabulary, 16k BPE merge operations and it is trained on unconstrained data.

PROMT BASIC TRANSFORMER uses separate vocabs (16k source + 32k target), and tied embeddings.

PROMT MULTILINGUAL 4-TO-EN is a multilingual system trained to translate from Croatian, Serbian, Slovak and Czech to English. It is a basic Transformer configuration with shared vocabulary.

PROMT MULTILINGUAL PL-EN is a Polish↔English system trained jointly in both directions. It uses basic Transformer configuration and shared vocabulary.

None of PROMT systems are constrained.

2.3.23 SJTU-NICT (Li et al., 2020)

SJTU-NICT represents two different main approaches. For News Translation Task, (1) cross-lingual language models (XLM) are used in an additional encoder to benefit from language-independent sentence representations from both the source and target side for Polish→English. For English→Chinese, which includes document-level information, three-stage training is used to train Longformer (Transformer with attention extended to the full document).

2.3.24 SRPOL (Krubiński et al., 2020)

No short description provided.

2.3.25 TALP UPC (Escolano et al., 2020)

No short description provided.

2.3.26 TENCENT TRANSLATION (Wu et al., 2020b)

No short description provided.

2.3.27 THUNLP (no associated paper)

No description provided.

2.3.28 TILDE (Krišlauks and Pinnis, 2020)

For WMT 2020, Tilde developed English↔Polish (separate constrained and unconstrained submissions) and Polish↔English (constrained only) NMT systems. Tilde experimented with morpheme splitting prior to byte-pair encoding, dual conditional cross-entropy filtering, sampling-based backtranslation of source-domain-adherent monolingual data, and right-to-left reranking. The submitted translations were produced using ensembles of Transformer base and Transformer big models, which were trained using back-translated data, and right-to-left re-ranking.

2.3.29 TOHOKU-AIP-NTT (Kiyono et al., 2020)

TOHOKU-AIP-NTT used Transformer-based Encoder-Decoder model with 8 layers and feed forward dimension of 8192. Synthetic data were created via beam back-translation from monolingual data available for each language and incorporated to the training using tagged backtranslation. The bitext was oversampled so that the model saw the bitext and synthetic data in 1:1 ratio. After training, the model was finetuned with newstest corpus.

An ensemble of four models was used to generate candidate translation, which were in turn re-ranked using scores from following components: (1) source-to-target right-to-left model, (2) target-to-source left-to-right model, (3) target-to-source right-to-left model, (4) masked language model (RoBERTa), and (5) uni-directional language model (Transformer-LM).

2.3.30 MULTILINGUAL-UBIQUS (Hernandez and Nguyen, 2020)

UBIQUS performed a single submission, based on an unconstrained multilingual setup. The approach consists of jointly training a traditional Transformer model on several agglutinative languages in order to benefit from them for the low-resource English-Inuktitut task. For that purpose,

the dataset was extended with other linguistically near languages (Finnish, Estonian), as well as in-house datasets introducing more diversity to the domain.

2.3.31 UEDIN

UEDIN **(Bawden et al., 2020a)** for the very low-resource English-Tamil involved exploring pretraining, using both language model objectives and translation using an unrelated high-resource language pair (German-English), and iterative backtranslation. For English-Inuktitut, UEDIN explored the use of multilingual systems.

UEDIN-DEEN **and** UEDIN-ENDE **(Germann, 2020)** ensemble big transformer models trained in three stages: First, base transformer models were trained on available high-quality parallel data. These models were used to rank and select parallel data from crawled and automatically matched parallel data (Paracrawl, Commoncrawl, etc.). 2nd-generation big transformers were then trained on the combined parallel data. These models were used for back-translation. Original and back-translated data was then used to the final 3rd-generation models.

2.3.32 UEDIN-CUNI (Germann et al., 2020)

UEDIN-CUNI CSEN STUDENT **and** UEDIN-CUNI ENCS STUDENT are compact, efficient student models that distill knowledge from larger teacher models. All models are variants of the transformer architecture. The teacher models were used to translate the source side of the training data to create synthetic training data for the student models.

2.3.33 UQAM_TANLE

No description provided.

2.3.34 VOLCTRANS (Wu et al., 2020a)

VOLCTRANS aims at building a general training framework which can be well applied to different translation. directions. Techniques used in the submitted systems include optional multilingual pre-training (mRASP) for low resource languages, very deep Transformer or dynamic convolution models up to 50 encoder layers, iterative back-translation, knowledge distillation, model ensemble and development set fine-tuning. The key ingredient of the process seems the strong focus on diversification of the (synthetic) training data, using multiple scalings of the Transformer model and dynamic convolution, random upsamplings of the parallel data, creation of multiple back-translated corpus variants or random ensembling which uses not a fixed set of ensembled models but rather a random checkpoint of each of them.

2.3.35 WECHAT (Meng et al., 2020)

WECHAT is based on the Transformer with effective variants and the DTMT architecture. The experiments include data selection, several synthetic data generation approaches (i.e., back-translation, knowledge distillation, and iterative in-domain knowledge transfer), advanced finetuning approaches and self-bleu based model ensemble.

2.3.36 WMTBIOMEDBASELINE (Bawden et al., 2020b)

WMTBIOMEDBASELINE are the baseline systems from the Biomedical Translation Task.

2.3.37 YOLO (no associated paper)

No description provided.

2.3.38 ZLABS-NLP

ZLABS-NLP used SentencePiece for subword segmentation, otherwise the model including hyper-parameters is the same as described by Ott et al. (2018) and implemented in FairSeq. Probably, OpenNMT-py was used during training (back-translation for Tamil).

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

Since running a comparison of *direct assessments* (DA, Graham et al., 2013, 2014, 2016) and relative ranking in 2016 (Bojar et al., 2016) and verifying a high correlation of system rankings for the two methods, as well as the advantages of DA, such as quality controlled crowd-sourcing and linear growth relative to numbers of submissions, we have employed DA as the primary mechanism for evaluating systems. 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.[5] 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) and multilingual surface realisation (Mille et al., 2018, 2019).

3.1.1 Source and Reference-based Evaluations

The earlier DA evaluations that we performed were all referenced based, as described above, however in 2018 we trialled source-based evaluation for the first time, in English to Czech translation. In this configuration, the human assessor is shown the source input and system output only (with no reference translation shown). This approach has the advantage of freeing up the human-generated reference translation so that it can be included in the evaluation to provide an estimate of human performance. As was the approach in WMT19, since we would like to restrict human assessors to only evaluate translation *into* their native language, we again restrict bilingual/source-based evaluation to evaluation of translation 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.

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), so we avoid it where possible. As detailed in Section 2, most of our test sets do not include reverse-created sentence pairs, except when there were resource constraints on the creation of the test sets.

3.1.3 Document Context

Prior to WMT19, the issue of including document context was raised within the community (Läubli et al., 2018; Toral et al., 2018) and at WMT19 a range of DA styles were subsequently tested that included document context. In WMT19, two options were run, firstly, an evaluation that included the document context "+DC" (with document context), and secondly, a variation that omitted document context "−DC" (without document context). This year, for language pairs for which document context was available in the test set, we therefore include this context when evaluating translations for systems. Although we include document context, 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 and 3.3 below for translation into English and out of English, resp.

In the following, we use the following abbreviations to describe annotation style: SR+DC for translation direction where assessors rank individual segments (Segment Ranking, SR) and have access to the full document, SR−DC for translation directions where document context is not available and assessors see individual sentences in random order.

Fully document-level evaluation (DR+DC, document-level ranking with document context available) as trialled last year where we asked for a single score given the whole document is problematic in terms of statistical power and inconclusive ties, as shown in Graham et al. (2019); Graham et al. (2020), and we subsequently did not include this approach for any into-English language this year.

As in previous years, the SR−DC annotation is organized into "HITs" (following the Mechanical Turk's term "human intelligence task"), each containing 100 screens.

[5] Past work has investigated the degree to which employment of a reference translation in DA evaluations could introduce bias into evaluation results and showed no significant evidence of reference-bias (Ma et al., 2017).

	Seg Rating + Doc Context (SR+DC)	Seg Rating − Doc Context (SR−DC)
Chinese to English	**M**	
Czech to English	**M**	
German to English	**M**	
Inuktitut to English		**M**
Khmer to English		**M**
Japanese to English	**M**	
Pashto to English		**M**
Polish to English	**M**	
Russian to English	**M**	
Tamil to English	**M**	

Table 7: Summary of human evaluation configurations for monolingual translation for into-English language pairs; M denotes reference-based/monolingual human evaluation in which the machine translation output was compared to human-generated reference

Language Pair	**Sys.**	**Assess.**	**Assess/Sys**
Czech→English	12	10,703	891.9
German→English	13	14,303	1,100.2
Inuktitut→English	11	13,897	1,263.4
Japanese→English	10	11,234	1,123.4
Khmer→English	7	7,944	1,134.9
Polish→English	14	14,146	1,010.4
Pashto→English	6	8,162	1,360.3
Russian→English	12	12,783	1,065.2
Tamil→English	14	8,899	635.6
Chinese→English	17	34,596	2,035.1
Total to-English	116	136,667	1,178.2

Table 8: Amount of data collected in the WMT20 manual evaluation campaign for evaluation into-English; after removal of quality control items.

3.2 Human Evaluation of Translation into-English

A summary of the human evaluation configurations run this year in the news task for into-English language pairs is provided in Table 7.

In terms of the News translation task manual evaluation for into-English language pairs, a total of 654 turker accounts were involved.[6] 654,583 translation assessment scores were submitted in total by the crowd, of which 166,868 were provided by workers who passed quality control.

System rankings are produced from a large set of human assessments of translations, each of which indicates the absolute quality of the output of a system. Table 8 shows total numbers of human assessments collected in WMT20 for into-English language pairs contributing to final scores for systems.[7]

[6]Numbers do not include the 2,233 workers on Mechanical Turk who did not pass quality control.

[7]Number of systems for WMT20 includes three "human" systems comprising human-generated reference translations used to provide human performance estimates.

3.2.1 Crowd Quality Control

We run two configurations of DA, one with document context, segment-rating with document context (SR+DC), for languages for which this information was available and one without document context, for the remainder, segment rating without document context (SR-DC). We describe quality control details and both methods of ranking systems for into-English language pairs in detail below.

Standard DA HIT Structure (SR−DC) In the standard DA HIT structure (without document context), three kinds of quality control translation pairs are employed as described in Table 9: 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.

In total, 60 items in a 100-translation HIT serve in quality control checks but 40 of those are regular judgments of MT system outputs (we exclude assessments of bad references and ordinary reference translations when calculating final scores). The effort wasted for the sake of quality control is thus 20%.

Also in the standard DA HIT structure, within each 100-translation HIT, the same proportion of translations are included from each participating system for that language pair. This ensures the final dataset for a given language pair contains roughly equivalent numbers of assessments for each participating system. This serves three purposes for making the evaluation fair. Firstly, for the point estimates used to rank systems to be reliable, a sufficient sample size is needed and the most efficient way to reach a sufficient sample size for all systems is to keep total numbers of judgments roughly equal as more and more judgments are collected. Secondly, it helps to make the evaluation fair because each system will suffer or benefit equally from an overly lenient/harsh human judge. Thirdly, despite DA judgments being absolute, it is known that judges "calibrate" the way they use the scale depending on the general observed translation quality. With each HIT including all participating systems, this effect is

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

averaged out. Furthermore apart from quality control items, HITs are constructed using translations sampled from the entire set of outputs for a given language pair.

Document-Level DA HIT Structure (SR+DC) Collection of segment-level ratings with document context (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.

Quality control items for this set-up was carried out as follows with the aim of constructing a HIT with as close as possible to 100 segments in total:

1. All documents produced by all systems are pooled;[8]

2. Documents are then sampled at random (without replacement) and assigned to the current HIT until the current HIT comprises no more than 70 segments in total;

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 in total within a HIT but without selecting documents in any systematic way such as selecting them based on fewest segments, for example.

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);

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 follows:

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 $\rfloor$

3.2.2 Annotator Agreement

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 fil-

[8] If a "human" system is included to provide a human performance estimate, it is also considered a system during quality control set-up.

ter 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-sourced worker who has not demonstrated that he/she can reliably score bad reference translations significantly lower than corresponding genuine system 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 10 shows the number of workers participating in the into-English translation evaluation who met our filtering requirement in WMT20 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.

3.3 Human Evaluation of Translation out-of-English

Human evaluation of out-of-English translations features a bilingual/source-based evaluation campaign that enlists the help of participants in the shared task. As usual, each team was asked to contribute around 8 hours annotation time, which we estimated at 16 HITs per each primary system submitted, with each HIT including 100 segment translations. Unfortunately, not all participating teams were able to provide requested number of assessments, hence, to collect the required number of assessments per MT system, we also employed external translators in a separate campaign. The contracted translators contributed with one third of total number of assessments. Both campaigns utilized document-level DA and were run for all out-of-English language pairs, which test sets include document-level segmentation.

For English→Khmer, English→Pashto, French→German, and German→French, whose test sets do not provide document boundaries, segment-level DA evaluation without document context (SR–DC) was performed, enlisting the effort of translators.

For English→Inuktitut, since we expected no participants to speak Inuktitut, the NRC hired native speakers through the Pirurvik Centre to conduct most of the DA evaluation. Due to the delays in starting the evaluation campaign, they were only able to complete the evaluation a few days before the conference, and could only annotate the news half of the test set. The Hansard half of the test set was not assessed in time for this report, but plans are being made to continue the evaluation after the conference. Updated rankings should be provided at a future date.

In terms of the News translation task document-level manual evaluation for out-of-English language pairs, a total of 1,189 researcher/translator accounts were involved, and 248,597 translation assessment scores were contributed in total (with quality control pairs), including 18,108 document ratings. For the segment-level campaigns (i.e. English→Khmer, English→Pashto, German→French and French→German) we had 300 accounts and 65872 scores collected in total. Statistics per language pair are summarized in Table 11. For data collection we again used the open-source Appraise[9] (Federmann, 2012). 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.

3.3.1 Document-Level Assessment

This year's human evaluation for out-of-English language pairs features an improved document-level direct assessment configuration that extends the context span to entire documents for a more reliable machine translation evaluation (Castilho

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

		All	(A) Sig. Diff. Bad Ref.	(A) & No Sig. Diff. Exact Rep.
SR−DC	Inuktitut→English	464	87 (19%)	81 (93%)
	Khmer→English	529	60 (11%)	56 (93%)
	Pashto→English	321	46 (14%)	46 (100%)
	Total	**1,126**	**169 (15%)**	**158 (93%)**
SR+DC	Czech→English	247	50 (20%)	43 (86%)
	German→English	343	84 (24%)	77 (92%)
	Japanese→English	422	81 (19%)	74 (91%)
	Polish→English	367	87 (24%)	77 (89%)
	Russian→English	360	109 (30%)	89 (82%)
	Tamil→English	235	71 (30%)	65 (92%)
	Chinese→English	878	178 (20%)	158 (89%)
	Total	**1,804**	**482 (27%)**	**423 (88%)**
	Overall	**2,930**	**651 (22%)**	**581 (90%)**

Table 10: 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.

Language Pair	Sys.	Assess.	Assess/Sys
English→Czech	13	37,535	2,887.3
English→German	17	19,102	1,123.6
English→Inuktitut	12	21,816	1,818.0
English→Japanese	12	24,341	2,028.4
English→Polish	15	20,162	1,344.1
English→Russian	10	21,618	2,161.8
English→Tamil	16	10,123	632.7
English→Chinese	14	46,207	3,300.5
Total document-level	109	200,904	1,843.2
German→French	7	14,470	2067.1
French→German	9	16,844	1871.6
English→Khmer	8	13,393	1,674.1
English→Pashto	7	13,267	1,895.3
Total segment-level	31	57,974	1,870.1

Table 11: Amount of data collected in the WMT20 manual document- and segment-level evaluation campaigns for bilingual/source-based evaluation out of English and non-English pairs.

et al., 2020; Laubli et al., 2020). It differs from SR+DC DA introduced in WMT19 (Bojar et al., 2019), and still used in into-English human evaluation this year, where a single segment from a document is provided on a screen at a time, followed by showing the entire document during annotation. Figure 6 shows a screenshot of the document-level direct assessment interface introduced this year.[10] Annotators see the entire document on a screen. 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.

3.3.2 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 minor modifications. Source-based DA allows to include human references in the evaluation as another system to provide an estimate of human performance. Human references were added to the pull of system outputs prior to sampling documents for tasks generation. If multiple references are available, which is the case for English→German (3 alternative reference translations, including 1 generated using the paraphrasing method of Freitag et al. (2020)) and English→Chinese (2 translations), each reference is assessed individually.

Since the annotations are made by researchers and professional translators who ensure a bet-

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

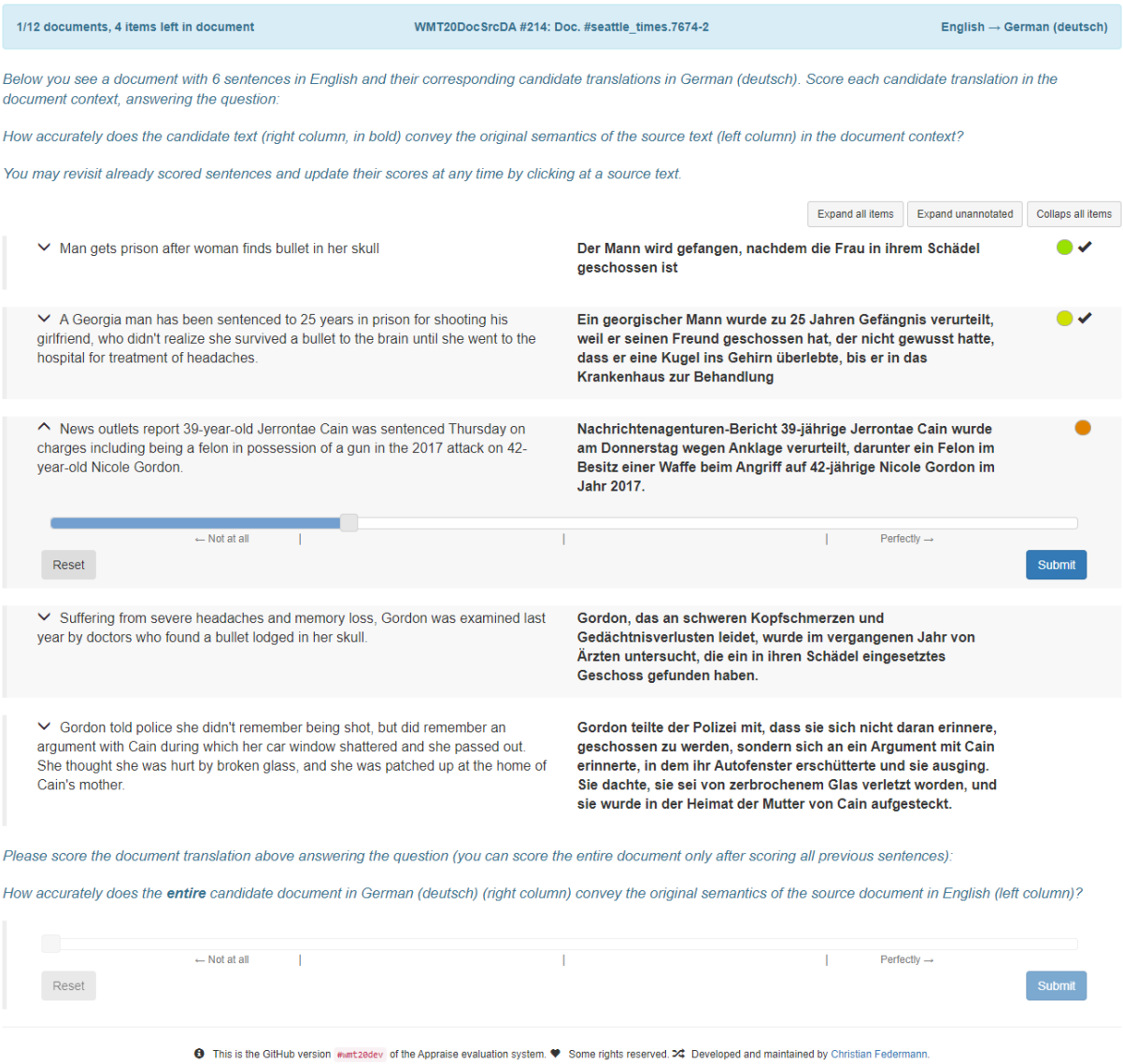

Figure 6: Screen shot of the new document-level DA configuration in the Appraise interface for an example assessment from the human evaluation campaign. The annotator is presented with the entire translated document randomly selected from competing systems (anonymized) and is asked to rate the translation of individual segments and the entire document on sliding scales.

ter quality of assessments than the crowd-sourced workers, only bad references are used as quality control items. Instead of sampling initial documents with close to 70 segments, we sample documents with 88 segments, and then a subset of documents with around 12 segments is selected to be converted into bad references. The remaining of the HIT creation process remains the same.

3.4 Producing the Human Ranking

In all set-ups, similar to previous years, system rankings were arrived at in the following way. Firstly, 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. This year all rankings for to-English translation were arrived at via segment ratings (SR−DC, SR+DC), 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 12). Results are also reported for average scores for systems, computed in the same way but without any score standardization applied (Ave % in Table 12).

Table 13 shows official news task results for translation out of English, where lines indicate

Chinese→English

Ave.	Ave. z	System
77.5	0.102	VolcTrans
77.6	0.089	DiDi-NLP
77.4	0.077	WeChat-AI
76.7	0.063	Tencent-Translation
77.8	0.060	Online-B
78.0	0.051	DeepMind
77.5	0.051	OPPO
76.5	0.028	THUNLP
76.0	0.016	SJTU-NICT
72.4	0.000	Huawei-TSC
76.1	−0.017	Online-A
74.8	−0.029	HUMAN
71.7	−0.071	Online-G
74.7	−0.078	dong-nmt
72.2	−0.106	zlabs-nlp
72.6	−0.135	Online-Z
67.3	−0.333	WMTBiomedBaseline

Czech→English

Ave.	Ave. z	System
78.3	0.118	CUNI-DocTransformer
77.5	0.071	OPPO
74.8	0.041	Online-B
75.3	0.034	CUNI-Transformer
73.8	0.018	Online-A
73.7	−0.037	SRPOL
74.1	−0.049	UEDIN-CUNI
74.1	−0.065	CUNI-T2T-2018
72.5	−0.069	Online-G
71.8	−0.080	Online-Z
71.9	−0.094	PROMT-NMT
72.0	−0.141	zlabs-nlp

German→English

Ave.	Ave. z	System
82.6	0.228	VolcTrans
84.6	0.220	OPPO
82.2	0.186	HUMAN
81.5	0.179	Tohoku-AIP-NTT
81.3	0.179	Online-A
81.5	0.172	Online-G
79.8	0.171	PROMT-NMT
82.1	0.167	Online-B
78.5	0.131	UEDIN
78.8	0.085	Online-Z
74.2	−0.079	WMTBiomedBaseline
71.1	−0.106	zlabs-nlp
20.5	−1.618	yolo

Inuktitut→English

Ave.	Ave. z	System
73.1	0.168	NiuTrans
72.9	0.167	Facebook-AI
71.2	0.100	CUNI-Transfer
70.7	0.096	Groningen
70.3	0.072	SRPOL
71.1	0.066	Helsinki
70.2	0.055	NRC
70.2	0.054	UEDIN
70.1	0.047	UQAM-TanLe
68.8	0.006	NICT-Kyoto
68.4	−0.035	OPPO

Japanese→English

Ave.	Ave. z	System
75.1	0.184	Tohoku-AIP-NTT
76.4	0.147	NiuTrans
74.1	0.088	OPPO
75.2	0.084	NICT-Kyoto
73.3	0.068	Online-B
70.9	0.026	Online-A
71.1	0.019	eTranslation
64.1	−0.208	zlabs-nlp
66.0	−0.220	Online-G
61.7	−0.240	Online-Z

Khmer→English

Ave.	Ave. z	System
69.0	0.168	Online-B
69.4	0.146	GTCOM
68.5	0.136	Huawei-TSC
62.6	−0.047	VolcTrans
58.1	−0.210	OPPO
56.9	−0.222	Online-Z
55.5	−0.282	Online-G

Pashto→English

Ave.	Ave. z	System
67.3	0.032	Online-B
66.7	0.024	GTCOM
65.5	−0.016	Huawei-TSC
62.7	−0.106	VolcTrans
62.1	−0.164	OPPO
61.0	−0.195	Online-Z

Polish→English

Ave.	Ave. z	System
77.2	0.131	SRPOL
76.7	0.097	Online-G
77.7	0.096	NICT-Rui
77.9	0.094	Online-B
78.1	0.085	SJTU-NICT
76.6	0.083	Online-A
75.2	0.050	OPPO
77.3	0.006	Online-Z
78.1	−0.003	CUNI-Transformer
76.1	−0.038	NICT-Kyoto
73.3	−0.041	VolcTrans
73.2	−0.048	PROMT-NMT
74.3	−0.072	Tilde
74.0	−0.130	zlabs-nlp

Russian→English

Ave.	Ave. z	System
79.3	0.124	Online-G
80.9	0.114	Online-A
79.7	0.113	OPPO
80.6	0.104	eTranslation
79.5	0.096	PROMT-NMT
80.2	0.072	Online-B
79.9	0.062	HUMAN
77.7	0.042	ariel xv
79.2	0.026	AFRL
76.0	−0.016	DiDi-NLP
75.2	−0.022	Online-Z
71.7	−0.153	zlabs-nlp

Tamil→English

Ave.	Ave. z	System
68.7	0.203	GTCOM
70.3	0.202	OPPO
68.9	0.176	Online-B
73.9	0.173	Facebook-AI
70.9	0.150	NiuTrans
71.9	0.116	VolcTrans
64.5	0.007	Online-Z
66.4	0.001	zlabs-nlp
67.5	−0.016	Microsoft-STC-India
60.8	−0.020	UEDIN
64.5	−0.068	Online-A
63.4	−0.078	DCU
53.7	−0.398	Online-G
53.9	−0.451	TALP-UPC

Table 12: Official results of WMT20 News Translation Task for translation into-English. Systems ordered by DA score z-score; systems within a cluster are considered tied; lines indicate clusters according to Wilcoxon rank-sum test $p < 0.05$; grayed entry indicates resources that fall outside the constraints provided.

clusters according to Wilcoxon rank-sum test $p < 0.05$. For evaluation of English→Inuktitut insufficient data resulted in a small sample size of human assessments per system and as a result some systems that fall within the same cluster are likely to do so simply due to low statistical power (Graham et al., 2020).

Human performance estimates arrived at by evaluation of human-produced reference translations are denoted by "HUMAN" in all tables. Note that "HUMAN-P" is a human-produced paraphrase of HUMAN-A, according to the method proposed by Freitag et al. (2020). 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.

Appendix A shows the underlying head-to-head significance test official results for all pairs of systems. All data collected during the human evaluation is available at `http://www.statmt.org/wmt20/results.html`.

In terms of human and machine quality comparisons in results, it is clear from the source-based evaluation of English to German and English to Chinese translation that human translators vary in performance, with each human translator represented in a distinct cluster. Without taking from the significant achievement of systems that have tied with a human translator, this fact should be taken into account when drawing conclusions about human parity. A tie with a single human translator should not be interpreted as a tie with human performance in general.

4 Test Suites

"Test Suites" have now become an established part of WMT News Translation. Their purpose is to complement the standard one-dimensional manual evaluation. Each test suite can focus on any aspect of translation quality and any subset of language pairs and MT systems.

Anyone can propose their own test suite and take part, and we also try to solicit evaluation from past successful test suite teams to support some cross-year insight.

Each team in the test suites track provides source texts (and optionally references) for any language pair that is being evaluated by WMT News Task. We shuffle these additional texts into the inputs of News Task and ship them as inputs to MT system developers jointly with the regular news texts. The shuffling happens at the document or sentence level as agreed with the test suite authors. (Shuffling at the level of sentences can lead to a very high number of documents in the final test set because each sentence is treated as a separate document.)

MT system developers may decide to skip these documents based on their ID but most of them process test suites along with the main news texts. After collecting the output translations from all WMT News Task Participants, test suites translations are made available back to the test suite authors for evaluation. Test suite sentences do not go through the manual evaluation as described in Section 3.

As in the previous years, test suites are not limited to the news domain, so News Task system may actually underperform on them.

4.1 Test Suite Details

The following paragraphs briefly describe each of the test suites. Please refer to the respective paper for all the details of the evaluation.

4.1.1 Covid Test Suite TICO-19

The TICO-19 test suite was developed to evaluate how well can MT systems handle the newly-emerged topic of COVID-19. Accurate automatic translation can play an important role in facilitating communication in order to protect at-risk populations and combat the *infodemic* of misinformation, as described by the World Health Organization. The test suite has no corresponding paper so its authors provided an analysis of the outcomes directly here.

The submitted systems were evaluated using the test set from the recently-released TICO-19 dataset (Anastasopoulos et al., 2020). The dataset provides manually created translations of COVID-19 related data. The test set consists of PubMed articles (678 sentences from 5 scientific articles), patient-medical professional conversations (104 sentences), as well as related Wikipedia articles (411 sentences), announcements (98 sentences from Wikisource), and news items (67 sentences from Wikinews), for a total of 2100 sentences.

Table 15 outlines the BLEU scores by each submitted system in the English-to-X directions, also breaking down the results per domain. The analysis shows that some systems are significantly more prepared to handle highly narrow-domain data. In

English→Chinese

Ave.	Ave. z	System
80.6	0.568	HUMAN-B
82.5	0.529	HUMAN-A
80.0	0.447	OPPO
79.0	0.420	Tencent-Translation
77.3	0.415	Huawei-TSC
77.4	0.404	NiuTrans
77.7	0.387	SJTU-NICT
76.6	0.373	VolcTrans
73.7	0.282	Online-B
73.0	0.241	Online-A
69.5	0.136	dong-nmt
68.5	0.135	Online-Z
70.1	0.122	Online-G
68.7	0.082	zlabs-nlp

English→Czech

Ave.	Ave. z	System
85.6	0.654	HUMAN
82.2	0.546	CUNI-DocTransformer
81.8	0.538	OPPO
80.8	0.505	SRPOL
80.5	0.458	CUNI-T2T-2018
80.4	0.441	eTranslation
79.3	0.434	CUNI-Transformer
77.1	0.322	UEDIN-CUNI
70.5	0.048	Online-B
69.1	0.017	Online-Z
68.7	0.008	Online-A
62.7	−0.216	Online-G
48.1	−0.760	zlabs-nlp

English→German

Ave.	Ave. z	System
90.5	0.569	HUMAN-B
87.4	0.495	OPPO
88.6	0.468	Tohoku-AIP-NTT
85.7	0.446	HUMAN-A
84.5	0.416	Online-B
84.3	0.385	Tencent-Translation
84.6	0.326	VolcTrans
85.3	0.322	Online-A
82.5	0.312	eTranslation
84.2	0.299	HUMAN-paraphrase
82.2	0.260	AFRL
81.0	0.251	UEDIN
79.3	0.247	PROMT-NMT
77.7	0.126	Online-Z
73.9	−0.120	Online-G
68.1	−0.278	zlabs-nlp
65.5	−0.338	WMTBiomedBaseline

English→Inuktitut (News only)

Ave.	Ave. z	System
90.5	0.574	HUMAN
75.3	0.425	MultiLingual-Ubiqus
77.4	0.409	CUNI-Transfer
71.9	0.369	NRC
74.6	0.368	Facebook-AI
79.2	0.364	NICT-Kyoto
71.6	0.339	Groningen
75.2	0.296	Helsinki
72.8	0.282	SRPOL
68.9	0.084	UQAM-TanLe
66.4	0.081	UEDIN
48.2	−0.384	OPPO

English→Japanese

Ave.	Ave. z	System
79.7	0.576	HUMAN
77.7	0.502	NiuTrans
76.1	0.496	Tohoku-AIP-NTT
75.8	0.496	OPPO
75.9	0.492	ENMT
71.8	0.375	NICT-Kyoto
71.3	0.349	Online-A
70.2	0.335	Online-B
63.9	0.159	zlabs-nlp
59.8	0.032	Online-Z
53.9	−0.132	SJTU-NICT
52.8	−0.164	Online-G

English→Polish

Ave.	Ave. z	System
88.6	0.672	HUMAN
76.4	0.493	SRPOL
75.6	0.435	eTranslation
76.3	0.383	VolcTrans
74.0	0.348	Tilde
70.6	0.316	Online-G
72.0	0.310	OPPO
72.4	0.299	NICT-Kyoto
69.7	0.272	Tilde
71.8	0.255	CUNI-Transformer
70.1	0.236	Online-B
69.0	0.219	SJTU-NICT
64.5	0.097	Online-A
63.9	−0.060	Online-Z
47.7	−0.538	zlabs-nlp

English→Russian

Ave.	Ave. z	System
91.8	0.681	HUMAN
81.5	0.469	Online-G
83.7	0.461	OPPO
79.6	0.404	ariel xv
80.3	0.336	Online-B
75.1	0.252	PROMT-NMT
76.2	0.222	DiDi-NLP
75.3	0.081	Online-A
71.3	0.035	zlabs-nlp
68.5	0.012	Online-Z

English→Tamil

Ave.	Ave. z	System
83.4	0.762	HUMAN
79.0	0.663	Facebook-AI
75.5	0.514	GTCOM
77.3	0.491	Online-B
77.4	0.480	OPPO
78.0	0.457	Online-A
76.7	0.424	VolcTrans
72.8	0.326	Online-Z
72.7	0.307	zlabs-nlp
72.2	0.296	Microsoft-STC-India
74.1	0.231	UEDIN
71.9	0.153	Groningen
68.1	−0.006	DCU
58.2	−0.407	TALP-UPC
53.8	−0.716	Online-G
49.6	−0.819	SJTU-NICT

English→Khmer

Ave.	Ave. z	System
77.4	0.478	GTCOM
76.1	0.435	Online-B
74.6	0.386	Huawei-TSC
73.3	0.349	HUMAN
71.1	0.266	VolcTrans
63.8	0.059	Online-Z
60.9	−0.061	OPPO
57.0	−0.164	Online-Z

English→Pashto

Ave.	Ave. z	System
73.0	0.244	GTCOM
71.9	0.180	Huawei-TSC
70.4	0.162	OPPO
69.7	0.158	Online-B
68.8	0.092	HUMAN
67.7	0.055	Online-Z
66.9	−0.029	VolcTrans

Table 13: Official results of WMT20 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 $p < 0.05$; grayed entry indicates resources that fall outside the constraints provided.

German→ French			French → German		
Ave.	Ave. z	System	Ave.	Ave. z	System
90.4	0.279	OPPO	89.8	0.334	VolcTrans
90.2	0.266	VolcTrans	89.7	0.333	OPPO
89.7	0.262	IIE	89.1	0.319	IIE
89.2	0.243	HUMAN	89.0	0.295	Online-B
89.1	0.226	Online-B	87.4	0.247	HUMAN
89.1	0.223	Online-A	87.3	0.240	Online-A
88.5	0.208	Online-G	87.1	0.221	SJTU-NICT
			86.8	0.195	Online-G
			85.6	0.155	Online-Z

Table 14: Official results of WMT20 News Translation Task for translation from French ↔ German. Systems ordered by DA score z-score; systems within a cluster are considered tied; lines indicate clusters according to Wilcoxon rank-sum test $p < 0.05$; grayed entry indicates resources that fall outside the constraints provided.

addition, the variance of the output quality across languages and across domains highlights the importance of building MT systems that can generalize across domains.

4.1.2 Document Coherence Check via Markable Annotation (Zouhar et al., 2020)

The test suite provided in 2020 by the ELITR project (Zouhar et al., 2020) follows upon Vojtěchová et al. (2019). The focus this year is on "markables", i.e. mainly domain-specific terms that have to be translated consistently and unambiguously throughout the whole document (except news where style may require variation) to maintain lexical coherence. Manual annotation of the translation of markables is contrasted with manual annotation of fluency and adequacy and also BLEU scores.

The test suite is limited to 4 English→Czech documents and 2 Czech→English documents, covering 215 markable occurrences across 4 different domains. The set of markables was collected in the first phase of the annotation, which amounted to 4k assessments across the systems. The second annotation phase with 6.5k assessments compared markable translations, always checking outputs of all the 13 competing MT systems but still considering the document-level context of each of them.

Among other things, the observations indicate that the better the system, the lower the variance in manual scores. Markables annotation then confirms that frequent errors like bad translation of a term need not be the most severe and conversely, even rare errors such as bad disambiguation, over-translation or disappearance of a term or its translation which conflicts with other terms in the document can be critical.

The comparison of MT outputs with the reference (hidden among MT systems) in the evaluation is also interesting. Man-made errors were always marked as less severe than those of MT. The annotation also suggests that one of the document-level systems outperformed the reference in markable evaluation if error severity and frequency are weighted equally.

Fluency and adequacy collected as average sentence-level scores (with access to the full documents of all systems) are curious, revealing perhaps more about the annotators than the MT systems.

4.1.3 Gender Coreference and Bias (Kocmi et al., 2020)

The test suite by Kocmi et al. (2020) focuses on the gender bias in professions (e.g. physician, teacher, secretary) for the translation from English into Czech, German, Polish and Russian. These nouns are ambiguous with respect to gender in English but exhibit gender in the examined target languages.

The test suite is based on the fact that a pronoun referring to the ambiguous noun can reveal the gender of the noun in the English source sentence. Once disambiguated, the gender needs to be preserved in translation. To correctly translate the given noun, the translation system thus has to correctly resolve the coreference link and transfer information from the pronoun to the noun in the

en→◌		Translation Accuracy by Domain (BLEU)				
	Overall	PubMed	Conv.	Wikisource	Wikinews	Wikipedia
Mandarin Chinese (zh)						
SJTU-NICT	57.83	68.88	41.49	33.57	55.97	53.45
OPPO	40.80	49.54	17.01	26.42	31.41	37.53
Online-B	39.55	53.92	23.22	26.09	34.13	32.65
Online-A	35.23	42.81	18.15	20.83	27.77	32.46
Online-G	33.14	38.08	13.06	20.80	26.28	31.74
zlabs-nlp	24.17	31.15	10.11	17.39	21.05	21.00
Online-Z	22.69	28.58	13.31	13.70	17.80	20.30
Khmer (km)						
Online-B	9.01	11.85	7.68	25.86	12.78	6.22
VolcTrans	7.57	12.93	2.35	21.11	4.30	4.63
Online-Z	7.29	9.08	3.38	20.94	5.27	5.65
OPPO	6.99	7.59	6.95	10.73	5.52	6.54
Online-G	2.72	3.10	3.60	1.13	1.70	2.59
Tamil (ta)						
Online-B	30.42	21.42	17.91	31.31	34.11	35.50
Facebook_AI	15.56	12.41	8.71	16.06	16.67	17.40
Online-A	14.49	12.03	7.85	14.78	13.93	16.00
OPPO	12.86	10.22	5.89	13.26	11.67	14.51
UEDIN	12.25	10.15	9.59	12.90	13.83	13.36
Microsoft_STC_India	11.91	9.48	6.49	12.07	12.56	13.33
Online-Z	11.70	9.45	10.87	13.52	10.10	12.96
VolcTrans	11.63	10.12	11.91	9.52	12.32	12.53
zlabs-nlp	10.32	8.91	5.85	9.64	10.90	11.20
DCU	9.70	7.66	7.79	8.44	9.36	10.91
Groningen	8.93	8.00	5.95	8.14	9.66	9.47
Online-G	7.32	6.79	8.42	8.32	5.59	7.58
TALP_UPC	6.25	5.77	3.48	5.47	7.32	6.54
SJTU-NICT	2.91	3.01	3.72	5.26	2.68	2.67
Pashto (ps)						
Online-B	36.56	49.26	26.94	12.15	8.85	32.25
VolcTrans	18.47	24.22	16.21	12.58	8.96	16.41
OPPO	18.24	21.88	13.98	14.40	7.98	17.15
Online-Z	15.14	18.59	13.57	12.87	7.60	13.93
Russian (ru)						
Online-B	40.20	29.71	26.37	22.90	40.44	46.38
Online-G	33.78	28.20	25.51	22.58	32.39	37.30
PROMT_NMT	32.69	27.45	24.82	21.90	30.39	36.05
ariel197197	32.40	25.44	28.33	22.17	37.04	35.96
OPPO	31.86	29.04	23.33	22.17	32.27	33.76
Online-A	29.84	24.76	21.13	20.53	27.54	33.07
zlabs-nlp	25.83	23.63	21.96	19.40	25.97	27.20
Online-Z	24.67	20.26	20.43	20.01	26.09	27.07

Table 15: TICO-19 test suite results on the English-to-X WMT20 translation directions.

antecedent (a less common direction of information flow), and then correctly express the noun in the target language. The success of the MT system in this test can be established automatically, whenever the gender of the target word can be automatically identified.

Kocmi et al. (2020) build upon the WinoMT (Stanovsky et al., 2019) test set, which provides exactly the necessary type of sentences containing an ambiguous profession noun and a personal pronoun which unambiguously (for the human eye) refers to it based the situation described. When extending WinMT with Czech and Polish, Stanovsky et al. have to disregard some test patterns but the principle remains.

The results indicate that *all* MT systems fail in this test, following gender bias (stereotypical patterns attributing the masculine gender to some professions and feminine gender to others) rather than the coreference link.

4.1.4 Linguistic Evaluation of German-to-English (Avramidis et al., 2020)

The test suite by DFKI covers 107 grammatical phenomena organized into 14 categories. Since 2018, the same set of phenomena are being tested annually (Macketanz et al., 2018; Avramidis et al., 2019).

Automatic evaluation is complemented with 45 hours of human annotation.

This year, the newcomers VOLCTRANS and TOHOKU-AIP-NTT perform particularly well in the tested phenomena, followed by the traditional systems UEDIN, ONLINE-B, ONLINE-G, and ONLINE-A.

The generally good news is that systems which participated in both WMT19 and WMT20 show an improvement this year. Given that the test suite target side remains undisclosed, these scores can be deemed absolute, unlike the official DA scores which are only relative within each year and set of systems.

The test suite allows to report these improvements per linguistic category and specifically for each MT system that participated in two consecutive years. The biggest improvements are observed in long distance dependencies or interrogatives, verb valency, ambiguity and punctuation, and we tend to attribute all these improvements to increased capacity (which allows increased sensitivity to long-range relations) of the models.

4.1.5 Word Sense Disambiguation (Scherrer et al., 2020b)

Scherrer et al. (2020b) is a followup of last year's evaluation (Raganato et al., 2019), assessing the ability of MT systems to disambiguate a word given its context of the sentence.

The underlying MuCoW (multilingual contrastive word sense disambiguation) dataset contains approximately 2k to 4k sentences per language pair selected from large parallel corpora to contain particularly ambiguous words.

This year, the focus was on language pairs that appeared both in WMT19 and WMT20 (and were available in the MuCoW dataset), namely English→Czech, English↔German, and English→Russian.

Comparing overall numbers across the years, Scherrer et al. (2020b) report that ambiguous words are correctly disambiguated in the majority of cases. Both precision (percentage of correct choices out of sentences where either good or bad expected translation was found) and recall (percentage of correct choices out of all sentences) are above 60 % and reaching 80 % for the best systems in a given language pair when mixing "in-domain" and "out-of-domain" evaluation. The "out-of-domain" synsets are those that are represented in the test suite with more than half of cases coming from the colloquial subtitle domain; other synsets are deemed "in-domain". The "in-domain" scores are generally higher, with precisions above 95 % for the best Czech and Russian systems. Across the years, no real improvement is however observed.

Three cases suggest that training systems at the level of documents decreases their performance in this sentence-level evaluation (each sentence forms a separate document): DocTransformer vs. Transformer by CUNI in 2019 and 2020 and Microsoft document-level vs. sentence-level submission in 2019.

5 Similar Language Translation

Most shared tasks at WMT (e.g. News, Biomedical) have historically dealt with translating texts from and to English. In recent years, we observed a growing interest in training systems to translate between languages other than English. This includes a number of papers applying MT to translate between pairs of closely-related languages, national language varieties, and dialects

of the same language (Zhang, 1998; Marujo et al., 2011; Hassani, 2017; Costa-jussà et al., 2018; Popović et al., 2020). To address this topic, the first Similar Language Translation (SLT) shared task at WMT 2019 has been organized. It featured data from three pairs of closely-related languages from different language families: Spanish - Portuguese (Romance languages), Czech - Polish (Slavic languages), and Hindi - Nepali (Indo-Aryan languages).

Following the success of the first SLT shared task at WMT 2019 and the interest of the community in this topic, we organize, for the second time at WMT, this shared task to evaluate the performance of state-of-the-art translation systems on translating between pairs of languages from the same language family. SLT 2020 features five pairs of similar languages from three different language families: Indo-Aryan, Romance, and South-Slavic. Translations were evaluated in both directions using automatic evaluation metrics presented in this section.

5.1 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 (Europarl v10, News Commentary v15, Wiki Titles v2), while some corpora were newly introduced (JRC Acquis). The released parallel HI–MR dataset was collected from news (Siripragada et al., 2020), PMIndia (Haddow and Kirefu, 2020) and Indic Wordnet (Bhattacharyya, 2010; Kunchukuttan, 2020a) datasets. All data were initially combined, tokenized using indic-nlp tokenizer (Kunchukuttan, 2020b) and randomly shuffled. From the combined corpus, we randomly extracted 49,434 sentences for the training set and the rest are used as development and test sets. For the South-Slavic language pairs we used large datasets available from Opus (Tiedemann and Nygaard, 2004)[11], more precisely the OpenSubtitles, MultiParaCrawl, DGT and JW300 data. Different to the other language groups, for monolingual data web corpora of the three languages (Ljubešić and Erjavec, 2011; Ljubešić and Klubička, 2014; Erjavec et al., 2015) were given to the participants.

Development and Test Data The development and test sets for Spanish–Catalan and Spanish–Portuguese language pairs were created from a corpus provided by Pangeanic[12]. First, we performed cleaning using CLEAN-CORPUS-N.PERL[13] script to retain sentences that have between 4 and 100 tokens. This narrowed the number of sentences to 1,287 and 1,535 in dev and test sets respectively. Finally, sentences containing metadata information were removed, which resulted in 1,283 and 1,495 sentences in dev and test sets respectively.

The aforementioned shuffled combined HI–MR dataset, 1411 sentences are used for development set and 3882 for the test set. Finally, the test set was equally split into two different test sets: 1941 sentences used for HI to MR and 1941 sentences were used for MR to HI.

For the Slovene–Croatian and Slovene–Serbian language pairs, development and test data were obtained from the Ciklopea translation agency[14] in form of a data donation from the Bisnode business intelligence company[15]. The data consists of public relations releases translated in various directions between the three languages. The data was cleaned, deduplicated and shuffled, resulting in 2,457 dev and 2,582 instances for the Slovene–Croatian pair, and 1,259 dev and 1,260 test instances in the Slovene–Serbian pair. Given that these translations sometimes form Slovene–Croatian–Serbian triangles, special care was invested in circumventing data leakage between development data on one side, and test data on the other, of the two language pairs.

5.2 Participants and Approaches

The second edition of the WMT SLT task attracted 68 teams who signed up to participate in the competition and 18 of them submitted their system outputs. In the end of the competition, 14 teams submitted system description papers which are referred to in this report. Table 22 summarizes the participation across language pairs and translation directions and includes references to the 14 system description papers.

Next we provide summaries for each of the entries we received:

A3-108 The team A3-108 submitted their system for HI–MR and MR–HI. The team initially

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

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

[13] https://github.com/moses-smt

[14] https://ciklopea.com

[15] https://www.bisnode.hr

Table 16: Corpora for the Hindi ↔ Marathi language pair.

		Corpus	Sentences
Parallel	Hindi ↔ Marathi	News	12,349
	Hindi ↔ Marathi	PM India	25,897
	Hindi ↔ Marathi	Indic WordNet	11,188
Monolingual	Hindi	News Crawl 2008-2019	32,609,161
	Hindi	IITB	45,075,242
	Hindi	hi.yyyy_nn.raw.xz 2012-2017	
	Marathi	News Crawl 2018-2019	326,748
	Marathi	mr.yyyy_nn.raw.xz 2012-2017	
Dev	Hindi ↔ Marathi		1,411
Test	Hindi ↔ Marathi		1,941

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

		Corpus	Sentences
Parallel	Spanish ↔ Catalan	Wiki Titles v2	446,326
	Spanish ↔ Catalan	DOGC v2	10,933,622
Monolingual	Spanish	Europarl v10	2,038,042
	Spanish	News Commentary v15	465,165
	Spanish	News Crawl 2007-2019	53,874,815
	Catalan	caWaC	24,745,986
Dev	Spanish ↔ Catalan		1,283
Test	Spanish ↔ Catalan		1,495

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

		Corpus	Sentences
Parallel	Spanish ↔ Portuguese	Europarl v10	1,801,845
	Spanish ↔ Portuguese	News Commentary v15	48,259
	Spanish ↔ Portuguese	Wiki Titles v2	649,833
	Spanish ↔ Portuguese	JRC-Acquis	1,650,126
Monolingual	Spanish	Europarl v10	2,038,042
	Spanish	News Commentary v15	465,165
	Spanish	News Crawl 2007-2019	53,874,815
	Portuguese	Europarl v10	2,016,635
	Portuguese	News Commentary v15	73,550
	Portuguese	News Crawl 2008-2019	9,392,574
Dev	Spanish ↔ Portuguese		1,283
Test	Spanish ↔ Portuguese		1,495

Table 19: Corpora for the Slovenian ↔ Croatian language pair.

		Corpus	Sentences
Parallel	Slovenian ↔ Croatian	OpenSubtitles v2018	15,636,933
	Slovenian ↔ Croatian	MultiParaCrawl v5	271,415
	Slovenian ↔ Croatian	JW300 v1	1,052,547
	Slovenian ↔ Croatian	DGT v2019	698,314
Monolingual	Slovenian	slWaC	46,251,729
	Croatian	hrWaC	64,577,734
Dev	Slovenian ↔ Croatian		2,457
Test	Slovenian ↔ Croatian		2,582

Table 20: Corpora for the Slovenian ↔ Serbian language pair.

		Corpus	Sentences
Parallel	Slovenian ↔ Serbian	OpenSubtitles v2018	16,426,054
Monolingual	Slovenian	slWaC	46,251,729
	Serbian	srWaC	24,073,253
Dev	Slovenian ↔ Serbian		1,259
Test	Slovenian ↔ Serbian		1,260

Table 21: Corpora for the Croatian ↔ Serbian language pair.

		Corpus	Sentences
Parallel	Croatian ↔ Serbian	SETimes	203,989

build SMT models for both language direction after three steps preprocessing: (i) default – indic_nlp_library[16] and moses tokenizer[17], (ii) morfessor[18] and (iii) BPE[19]. These SMT models were used for back-translation. Finally, these back-translation data were used to train their NMT system.

ADAPT-DCU The ADAPT-DCU team participated in the SLT task on the Croatian–Slovene and Serbian–Slovene language pairs. The team's submissions were based on the Sockeye implementations of the Transformer, with a joint 32k-large BPE vocabulary for all three languages. The submission were regularly multilingual (having Slovene on one side and Croatian and Serbian on the other). The team used only OpenSubtitles bilingual training data, considering other available data to be too noisy. The basic implementation of the multilingual system was submitted as the second contrastive system, the multilingual implementation trained on filtered parallel data as the first contrastive system, while the primary submission included backtranslation of target monolingual data of segments similar to the development data. By performing n-gram-character-based filtering of training data the training time was cut in half with a minor improvement on the translation quality, while the largest improvements in translation quality were obtained by back-translating data similar to development data (between 8 and 14 BLEU points).

f1plusf6 During preprocessing as Marathi and Hindi are rich in terms of morphology, Applied two way segmentation as preprocessing, first supervised and unsupervised word based morphological segmentation and then BPE based segmentation to tackle low-resource language pairs. The participants used shared vocab across training and utilised POS based features on the source side to create initial models for both directions.

For preparing unsupervised back-translation parallel data they used aligned embedding space to generate word by word parallel sentences for both language directions. They also prepared initial models from the provided parallel data for back translation from monolingual data and pruned back-translation pairs based on perplexity score. Their model is based on Luong's attention on bi-LSTM network, copy attention on dynamically generated dictionary with label smoothing and dropouts to reduce overfitting.

Fast-MT Fast-MT team submitted their NMT system where Transformers and Recurrent Attention models are effectively used. They combined the recurrence based layered encoder-decoder model with the Transformer model. Their submitted system for Indo-Aryan Language (Hindi to Marathi) pair is trained on the parallel corpus of the training dataset provided by the organizers.

IIAI IIAI TEAM participate in both directions of the Hindi–Marathi translation task. Their primary submission is a transformer model trained on the released parallel and back-translated monolingual data. The team jointly learned BPE from the merged source–target corpus. After BPE, sentences were corrupted and reconstructed using the two ways:(i) 15% of the subwords in the sentence are randomly selected and masked, (ii) 15% of the subwords are randomly selected one by one and swapped with another randomly-selected subword in the sequence.

[16] https://anoopkunchukuttan.github.io/indic_nlp_library/
[17] https://github.com/moses-smt/mosesdecoder
[18] https://github.com/aalto-speech/morfessor
[19] https://github.com/rsennrich/subword-nmt

Team	System Description Paper
A3108	Yadav and Shrivastava (2020)
ADAPT-DCU	Popović and Poncelas (2020)
f1plusf6	Mujadia and Sharmaa (2020)
FAST-MT	Dhanani and Rafi (2020)
IIAI	
IIT-DELHI	Madaan et al. (2020)
INFOSYS	Rathinasamy et al. (2020)
IPN-CIC	Menéndez-Salazar et al. (2020)
NICT	
NITS-CNLP	Laskar et al. (2020)
NLPRL	Kumar et al. (2020)
NLPRL-IITBHU	
NUIG-Panlingua-KMI	Ojha et al. (2020)
NUST_FJWU	Haq et al. (2020)
Prompsit	
UBC-NLP	Adebara et al. (2020)
UPCTALP	Boncompte and Costa-jussà (2020)
WIPRO-RIT	Pal and Zampieri (2020)

Table 22: The teams that participated in the SLT 2020 task and their system description papers.

IITDELHI Team IITDELHI participated in the SLT task on Hindi–Marathi and Spanish–Portuguese language pairs. The team's primary submission builds on fine-tuning over pretrained mBART. They used pre-trained weights of the mBART model (Liu et al., 2019), which is pre-trained on large amounts of monolingual data for 25 languages including Spanish, however Portuguese is not there. The authors initialized a Transformer architecture with 12 encoder and decoder layers using the pre-trained weights, and then directly fine-tuned with the released training data. The authors conclude that mBART is helpful for transfer learning, even though the languages that are not available in the pre-trained model.

INFOSYS Infosys system for Hindi–Marathi (Primary) task is designed to learn the nuances of translation of this low resource language pair by taking advantage of the fact that the source (Hindi) and target (Marathi) languages are same alphabet languages. This system is an ensemble of FairSeq model built on anonymized parallel data and FairSeq back-translation model. The common words/tokens between source and target languages are anonymized during pre-processing upon which the FairSeq model is trained. The input statements during inferencing are anonymized based on the vocabulary of common tokens prepared during training and the predicted statements are de-anonymized during post-processing accordingly. This improved the accuracy (BLEU) of FairSeq considerably. Pre-processing also applies traditional parallel corpus filtering techniques to clean parallel data followed by domain specific techniques. There were records containing multiple statements delimited by slashes, where the domain specific techniques are applied to transform them in to records that retain only the matching single statement, identified based on its syntactic similarity with its parallel statement. Synthetic data generated with the mono-lingual (Marathi) data during FairSeq back-translation has unknown words (w.r.t parallel data vocabulary), resulting unknown words during prediction, which are downvoted while ensembling.

IPN-CIC This team participated in the Spanish-Portuguese language pair. The systems used the Transformer architecture with a fine-tuning for domain adaptation. The team proposed experiments on the kind of tokens used (words and sub-word units) and the initialization of the word embeddings in the systems using either a random initialization or pre-trained word embeddings.

NICT NICT participated in two language pairs: Hindi–Marathi and Spanish–Catalan, for both translation directions. Their primary submission is an unsupervised NMT system, initialized with a pre-trained cross-lingual language model (XLM), that has been trained using only the monolingual data provided by the organizers. They used the

standard hyper-parameters for training XLM and unsupervised NMT. Their contrastive submission is the same but supervised NMT system trained on the combination of the released bilingual and monolingual data.

NITS-CNLP NITS-CNLP system for HI–MR and MR–HI translation is based on cross-lingual language modelling with masked language modeling and translation language modeling. These language models were pre-trained on monolingual corpus and fine-tuned on parallel data following the architecture of Conneau and Lample (2019) and employing 6 layers with 8 attention heads and with 32 batch size, trained on a single GPU.

NLPRL This system submitted by the NLPRL team for the HI-MR is based on the Transformer approach. The system were trained on only the released parallel corpus. The team used SentencePiece library for preprocessing and set vocabulary size of 5000 symbols for source and target byte-pair encoding, respectively.

NLPRL-BHU The team participated in the HI $\leftrightarrow$ MR language pair. The participants used byte pair encoding to preprocess the data and fairseq library with the GRU-transformer for training.

NUIG-Panlingua-KMI The NUIG-Panlingua-KMI team explored phrase-based SMT, dependecy-based SMT method and neural method (used subword) for Hindi$\leftrightarrow$Marathi language pair.

NUST-FJWU NUST-FJWU system is an extension of state-of-the-art Transformer model with hierarchical attention networks to incorporate contextual information. During training the model used back-translation.

Prompsit This team is participating with a rule-based system based on Apertium (Forcada et al., 2009-11). Apertium is a free/open-source platform for developing rule-based machine translation systems and language technology that was first released in 2005. Apertium is hosted in Github where both language data and code are licensed under the GNU GPL. It is a research and business platform with a very active community that loves small languages. Language pairs are at a very different level of development and output quality in the platform, depending on two main variables: how much funded or in-kind effort has been devoted to it and the nature of the languages itself (the closer, the better).

UBC-NLP The UBC-NLP team participated in the SLT task on all the available language pairs. The team regularly used all the parallel data and trained 6-layer Transformer models based on the Fairseq library. Only for the Slovene–Croatian language pair the team performed back-translation, noticing a 3 BLEU point improvement in the results. This team obtained better results with bilingual than with multilingual models (training a single model for all language groups).

UPCTALP The UPCTALP participated in the Romance pairs. This team made use of the Transformer architecture improved with multilingual, back-translation and fine-tuning techniques. Each of this techniques improved over the previous one.

WIPRO-RIT WIPRO-RIT submitted their system to the SLT 2020 Indo-Aryan track. The presented system is a single multilingual NMT system based on the transformer architecture that can translate between multiple languages. The presented model is inspired from the model described in Johnson et al. (2017). WIPRO-RIT achieved competitive performance ranking 1st in Marathi to Hindi and 2nd in Hindi to Marathi translation among 22 systems.in Hindi to Marathi translation among 22 systems.

5.3 Results

We present results for the three language families: three different language families: Indo-Aryan (Hindi - Marathi), Romance (Spanish - Catalan, Spanish - Portuguese), and South-Slavic (Slovene - Croatian, Slovene - Serbian), all of them in the two possible directions. Like last year edision, the second edition of the Similar Translation Task evaluation was also performed on automatic basis using BLEU (Papineni et al., 2002), RIBES (Isozaki et al., 2010) and TER (Snover et al., 2006) measures. Each language direction is reported in one different table which contain information of the team; type of system, either contrastive (CONTRASTIVE) or primary (PRIMARY), and the BLEU, RIBES and TER results. The scores are sorted by BLEU. In general, primary systems tend to be better than contrastive systems, as expected, but there are some exceptions.

This year we recived major number of participants for the case of Indo-Aryan language group

i.e. Hindi–Marathi (in both directions). We received 22 submissions from 14 teams. The best systems (INFOSYS) based on BLEU for Hindi–Marathi achieved score 18.26, however based on other evaluation matric WIPRO-RIT achieved the best 62.45 RIBES and around 72 TER (see Table 23). While in the other direction Marathi–Hindi the best performing system (WIPRO-RIT) reached 24.53 of BLEU and 66.39 on TER, but based on RIBES score 66.83, IITDELHI performed the best (see Table 24).

Similarly to the previous edition of the SLT shared task, participants could submit systems for the Spanish–Portuguese language pair (in both directions). The best systems for Spanish-to-Portuguese achieved over 32 BLEU and around 52 TER. While in the opposite direction (Portuguese-to-Spanish) the best performing system reached 33.82 of BLEU, but its TER score was 52.41, which is higher than in the case of best performing Spanish-to-Portuguese systems. 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 86 BLEU and below 8 TER. In the case of Catalan-to-Spanish, the best systems scored around 77 BLEU and less than 15 TER.

A new language group in this year's SLT task is the group of (Western) South Slavic languages - Slovene, Croatian and Serbian, forming two language pairs - Slovene–Croatian and Slovene–Serbian, with one additional twist given the very high mutual intelligibility of Croatian and Serbian. The best systems for Slovene-to-Croatian achieved 36 BLEU and 43 TER, which is significantly worse than the results of the same best-performing system in the opposite direction - 43 BLEU and 36 TER. On the Slovene–Serbian pair a similar phenomenon can be observed - Slovene to Serbian achieving 39 BLEU and 40 TER, while the opposite direction achieves 47 BLEU and 33 TER. The reason for such a significant lack of symmetry is the better performance of the systems translating into Slovene, probably given that (Croatian and Serbian) multi-source translation (into Slovene) is simpler than multi-target translation, which was, finally, propagated to the back-translation procedure, increasing the difference between the directions even further.

5.4 Summary

In this section, we presented the results of the WMT SLT 2020 task. The second iteration of this competition featured data from five language pairs from three different language families: Hindi-Marathi; Spanish-Catalan and Spanish-Portuguese; Sloven-Croatian and Slovene-Serbian. We evaluated the systems translating in both directions of the language pair using three automatic metrics: BLEU, RIBES, and TER. We observed that the performance varies widely between language pairs. For example. the best performing systems trained to translate between Catalan and Spanish in both directions obtained significantly higher results than those trained to translate between other language pairs.

In terms of participation, SLT received system submissions from 18 teams. In the end of the competition, 14 teams wrote system description papers that appear in the WMT proceedings. The list of teams with references to the respective system description paper is presented in Table 22. Finally, short summaries of each entry, based on the description provided by the participants, were also presented in this section.

6 Conclusion

This paper presented the results of WMT20 news translation and similar language translation shared tasks, as well as the extra test suites added to the news translation task. Our main findings rank participating systems in their sentence-level and document-level translation quality, as assessed in a large-scale manual evaluation using the method of Direct Assessment (DA).

For out-of-English language pairs, DA was modified so that the context of the whole document is available while judging individual sentences and assessors are allowed to return to any sentence judgement within the document.

As in previous years, the effect of translationese (translating from a source which itself was produced in translation) was avoided except lower-resourced Inuktitut↔English, Pashto↔English, Khmer↔English, and German↔French by creating reference translations always in the same direction as the MT systems are run. Furthermore, 8 out-of-English language pairs would not need human reference for our evaluation at all because the assessors are evaluating translation candidates bilingually, comparing them to the source

Team	Type	BLEU ↑	RIBES ↑	TER ↓
INFOSYS	PRIMARY	18.26	56.73	76.48
WIPRO-RIT	PRIMARY	16.62	62.45	72.23
WIPRO-RIT	CONTRASTIVE2	15.42	61.02	73.59
IITDELHI	PRIMARY	15.14	61.06	74.63
IIAI	CONTRASTIVE	14.99	52.11	85.77
IITDELHI	CONTRASTIVE	14.91	57.63	81.19
IIAI	PRIMARY	14.73	52.80	86.13
WIPRO-RIT	CONTRASTIVE1	13.25	58.51	76.17
NLPRL	PRIMARY	12.50	58.66	76.86
NITS-CNLP	PRIMARY	11.59	57.76	79.07
A3108	PRIMARY	11.41	57.20	79.96
A3108	CONTRASTIVE	10.21	55.17	82.01
NUIG-Panlingua-KMI	CONTRASTIVE	9.76	52.18	91.49
NUIG-Panlingua-KMI	PRIMARY	9.38	51.88	91.24
f1plusf6	PRIMARY	5.49	43.74	94.60
f1plusf6	CONTRASTIVE	5.41	43.49	94.52
FAST-MT	PRIMARY	3.68	31.14	97.64
NICT	CONTRASTIVE	3.41	42.43	-
NICT	PRIMARY	1.26	31.20	-
UBC-NLP	PRIMARY	0	1	-
UBC-NLP	CONTRASTIVE	0	0.12	-

Table 23: Results for Hindi to Marathi translation.

Team	Type	BLEU ↑	RIBES ↑	TER ↓
WIPRO-RIT	PRIMARY	24.53	66.23	66.39
IITDELHI	PRIMARY	24.53	66.83	67.25
WIPRO-RIT	CONTRASTIVE2	22.93	65.89	68.11
WIPRO-RIT	CONTRASTIVE1	22.69	65.01	68.13
A3108	CONTRASTIVE	21.11	60.76	77.28
NLPRL	PRIMARY	20.72	64.46	71.04
IIAI	CONTRASTIVE	20.32	59.56	79.32
IIAI	PRIMARY	20.04	58.95	80.27
IITDELHI	CONTRASTIVE	18.74	58.56	77.22
A3108	PRIMARY	18.32	59.31	77.35
f1plusf6	PRIMARY	18.14	60.86	78.27
NUIG-Panlingua-KMI	CONTRASTIVE	17.39	58.84	81.15
NUIG-Panlingua-KMI	PRIMARY	17.38	59.31	81.47
f1plusf6	CONTRASTIVE	17.17	60.69	78.18
NITS-CNLP	PRIMARY	15.44	61.13	75.96
NICT	CONTRASTIVE	11.20	56.13	-
FAST-MT	PRIMARY	9.02	46.96	88.68
NUST_FJWU	CONTRASTIVE	6.79	46.27	91.28
NUST_FJWU	PRIMARY	6.71	43.19	93.74
NICT	PRIMARY	6.28	50.14	-
NLPRL-IITBHU	PRIMARY	0.12	7.66	-
UBC-NLP	PRIMARY	0.09	7.19	-
UBC-NLP	CONTRASTIVE	0	0.09	-

Table 24: Results for Marathi to Hindi translation.

Team	Type	BLEU ↑	RIBES ↑	TER ↓
Prompsit	PRIMARY	77.08	95.71	12.35
NICT	CONTRASTIVE	76.67	93.33	14.22
UPCTALP	PRIMARY	68.84	89.83	20.09
NICT	PRIMARY	68.43	92.13	19.47
UBC-NLP	PRIMARY	0.17	4.81	-
UBC-NLP	CONTRASTIVE	0	1.50	-

Table 25: Results for Catalan to Spanish translation.

Team	Type	BLEU ↑	RIBES ↑	TER ↓
Prompsit	PRIMARY	86.48	97.37	7.716
Prompsit	CONTRASTIVE	81.36	96.64	10.15
UPCTALP	PRIMARY	60.50	90.25	25.80
NICT	CONTRASTIVE	59.05	90.73	25.90
NICT	PRIMARY	51.97	88.30	31.68
UBC-NLP	CONTRASTIVE	9.53	64.17	77.42
UBC-NLP	PRIMARY	8.49	58.93	84.16

Table 26: Results for Spanish to Catalan translation.

Team	Type	BLEU ↑	RIBES ↑	TER ↓
UPCTALP	PRIMARY	33.82	76.04	52.41
IITDELHI	PRIMARY	32.84	74.84	52.65
Prompsit	PRIMARY	30.27	75.37	54.46
IPN-CIC	PRIMARY	28.38	72.24	56.27
IPN-CIC	CONTRASTIVE1	27.98	72.11	56.16
IPN-CIC	CONTRASTIVE2	27.41	75.18	57.28
UBC-NLP	CONTRASTIVE	0.06	1.50	-
UBC-NLP	PRIMARY	0	5.86	-

Table 27: Results for Portuguese to Spanish translation.

Team	Type	BLEU ↑	RIBES ↑	TER ↓
IIT-DELHI	PRIMARY	32.69	74.05	51.74
UPCTALP	PRIMARY	32.33	73.04	52.06
IPN-CIC	PRIMARY	27.08	72.98	55.34
Prompsit	PRIMARY	26.91	75.79	54.63
Prompsit	CONTRASTIVE	26.81	75.71	54.73
IPN-CIC	CONTRASTIVE1	23.91	71.55	57.55
IPN-CIC	CONTRASTIVE2	23.90	73.73	58.07
UBC-NLP	PRIMARY	17.06	52.55	76.21
UBC-NLP	CONTRASTIVE	4.47	52.72	88.13

Table 28: Results for Spanish to Portuguese translation.

Team	Type	BLEU ↑	RIBES ↑	TER ↓
ADAPT-DCU	PRIMARY	43.41	73.77	35.8
ADAPT-DCU	CONTRASTIVE2	29.04	68.71	48.74
ADAPT-DCU	CONTRASTIVE1	26.96	64.02	50.73
UBC-NLP	PRIMARY	0.07	1.03	-
UBC-NLP	CONTRASTIVE	0	0.25	-

Table 29: Results for Croatian to Slovene translation.

Team	Type	BLEU ↑	RIBES ↑	TER ↓
ADAPT-DCU	PRIMARY	35.56	72.04	43.19
ADAPT-DCU	CONTRASTIVE1	27.63	70.53	49.91
ADAPT-DCU	CONTRASTIVE2	23.3	60.8	52.79
UBC-NLP	PRIMARY	22.26	64.41	64.5
UBC-NLP	CONTRASTIVE	1.68	35.35	-

Table 30: Results for Slovene to Croatian translation.

Team	Type	BLEU ↑	RIBES ↑	TER ↓
ADAPT-DCU	PRIMARY	47.45	75.11	32.61
ADAPT-DCU	CONTRASTIVE1	33.5	70.86	44.58
ADAPT-DCU	CONTRASTIVE2	30.28	65.92	47.77
UBC-NLP	CONTRASTIVE	0	0.39	-
UBC-NLP	PRIMARY	0	1.3	-

Table 31: Results for Serbian to Slovene translation.

Team	Type	BLEU ↑	RIBES ↑	TER ↓
ADAPT-DCU	PRIMARY	39.16	73.37	39.81
ADAPT-DCU	CONTRASTIVE1	29.79	70.24	47.55
ADAPT-DCU	CONTRASTIVE2	25.7	64.81	50.51
UBC-NLP	PRIMARY	20.18	63.37	65.56
UBC-NLP	CONTRASTIVE	2.01	38.87	-

Table 32: Results for Slovene to Serbian translation.

text (as opposed to the reference) in these language pairs. The reference translations are nevertheless included as evaluation, hidden among participating MT systems.

This year, English→German included two independent reference translations and one human-produced paraphrase, and English→Chinese included two references. Each of these translations ended up significantly differing in quality from the other ones. In German↔English and also Chinese→English and English→Inuktitut, some MT systems fall in the same cluster with human translation. The observed variance of human translation quality however demands modesty before making any claims about human parity.

The need for cautious interpretation of the results is also strengthened by the fact that even in English→German and English→Czech where human translation was seemingly significantly surpassed in 2018 and/or 2019, the result is not confirmed this year. Furthermore and similarly to previous year, a test suite this year again suggests that some aspects of translation are not handled by current systems at all. This year all MT systems fall into the gender bias trap (Kocmi et al., 2020) and they tend to make more severe errors than humans (Zouhar et al., 2020).

The results of the task on similar language translation indicate that the performance when translating between pairs of closely-related languages is extremely varied across different language pairs. The best performing systems trained to translate between Catalan and Spanish, for example, obtained significantly higher results in both directions than those trained to translate between other language pairs in terms of BLEU, RIBES, and TER.

Acknowledgments

Translation of the test sets for the News task was sponsored by the EU H2020 projects ELITR and Bergamot (English-Czech), and GoURMET (English-Tamil), by Yandex (Russian-English), Microsoft (Chinese-English and German-English), Tilde (Polish-English), Lingua Custodia (French-German), Facebook (Pashto-English, Khmer-English), the University of Tokyo and NTT (Japanese-English). The human evaluation was co-funded by Google and Microsoft. For the human evaluation of English→Inuktitut, we are grateful for the funding received from the NRC and to the workers of the Pirurvik Centre who did the evaluation. We are also grateful to the many workers who contributed to the

human evaluation via Mechanical Turk. We would like to thank Roland Kuhn for advising on the Inuktitut↔English task organization and the Nunavut Maligaliurvia (Legislative Assembly of Nunavut) and Nunatsiaq News for supplying all the Inuktitut↔English parallel data.

The organizers of the similar languages task would like to thank Ciklopea and Bisnode for the Croatian, Serbian, and Slovene data, and Pangeanic for the Catalan, Portuguese, and Spanish data. The work of the organizers of the similar languages task is supported in part by the Spanish Ministerio de Ciencia e Innovación, through the postdoctoral senior grant Ramón y Cajal and by the Agencia Estatal de Investigación through the projects EUR2019-103819, PCIN-2017-079 and PID2019-107579RB-I00 / AEI / 10.13039/501100011033.

This work was supported in part 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.

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.

References

Ife Adebara, El Moatez Billah Nagoudi, and Muhammad Abdul Mageed. 2020. Translating similar languages: Role of mutual intelligibility in multilingual transformers. In *Proceedings of WMT*.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federman, Dmitriy Genzel, Francisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. Tico-19: the translation initiative for covid-19. In *NLP COVID-19 Workshop*, Online.

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, 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.

Loïc Barrault, Magdalena Biesialska, Marta R. Costa-jussà, Fethi Bougares, and Olivier Galibert. 2020. Findings of the first shared task on lifelong learning machine translation. 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.

Rachel Bawden, Alexandra Birch, Radina Dobreva, Arturo Oncevay, Antonio Valerio Miceli Barone, and Philip Williams. 2020a. The university of edinburgh's english-tamil and english-inuktitut submissions to the wmt20 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Rachel Bawden, Giorgio Maria Di Nunzio, Cristian Grozea, Inigo Jauregi Unanue, Antonio Jimeno Yepes, Nancy Mah, David Martinez, Aurélie Névéol, Mariana Neves, Maite Oronoz, Olatz Perez-de Viñaspre, Massimo Piccardi, Roland Roller, Amy Siu, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Dina Wiemann, and Lana Yeganova. 2020b. Findings of the wmt 2020 biomedical translation shared task: Basque, italian and russian as new additional languages. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Chao Bei, Hao Zong, Qingmin Liu, and Conghu Yuan. 2020. Gtcom neural machine translation systems for wmt20. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Pushpak Bhattacharyya. 2010. IndoWordNet. In *Proceedings of LREC*.

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. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

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.

Pere Vergés Boncompte and Marta R. Costa-jussà. 2020. Multilingual neural machine translation: Case-study for catalan, spanish and portuguese romance languages. In *Proceedings of WMT*.

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, 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, Christian Girardi, and Marcello Federico. 2012. WIT3: web inventory of transcribed and translated talks. In *Proc. of EAMT*, pages 261–268.

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*, Online. Association for Computational Linguistics.

Peng-Jen Chen, Ann Lee, Changhan Wang, Naman Goyal, Angela Fan, Mary Williamson, and Jiatao Gu. 2020a. Facebook ai's wmt20 news translation task submission. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Tanfang Chen, Weiwei Wang, Wenyang Wei, Xing Shi, Xiangang Li, Jieping Ye, and Kevin Knight. 2020b. The didi machine translation system for wmt20. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7059–7069. Curran Associates, Inc.

Marta R. Costa-jussà, Roger Creus, Oriol Domingo, Albert Domínguez, Miquel Escobar, Cayetana López, Marina Garcia, and Margarita Geleta. 2020. MT-Adapted Datasheets for Datasets: Template and Repository. *arXiv e-prints*, page arXiv:2005.13156.

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

Farhan Dhanani and Muhammad Rafi. 2020. Attention transformer model for translation of similar languages. In *Proceedings of WMT*.

Prajit Dhar, Arianna Bisazza, and Gertjan van Noord. 2020. Linguistically motivated subwords improve english-tamil translation: University of groningen's submission to wmt-2020. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Tomaž Erjavec, Nikola Ljubešić, and Nataša Logar. 2015. The slwac corpus of the sloveneweb. *Informatica*, 39(1).

Carlos Escolano, Marta R. Costa-jussà, and José A. R. Fonollosa. 2020. The talp-upc machine translation systems for wmt20 news translation task: Multilingual adaptation for low resource mt. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Miquel Esplà-Gomis. 2009. Bitextor: a free/open-source software to harvest translation memories from multilingual websites. In *MT Summit Workshop on New Tools for Translators*. International Association for Machine Translation.

M. Amin Farajian, António V. Lopes, André F. T. Martins, Sameen Maruf, and Gholamreza Haffari. 2020. Findings of the wmt 2020 shared task on chat translation. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

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

Mikel L. Forcada, Francis M. Tyers, and Gema Ramírez Sánchez. 2009-11. The apertium machine translation platform: five years on.

Alexander Fraser. 2020. The wmt 2020 shared tasks in unsupervised mt and very low resource supervised mt. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

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

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Ulrich Germann. 2020. The university of edinburgh's submission to the german-to-english and english-to-german tracks in the wmt 2020 news translation and zero-shot translation robustness tasks. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Ulrich Germann, Roman Grundkiewicz, Martin Popel, Radina Dobreva, Nikolay Bogoychev, and Kenneth Heafield. 2020. Speed-optimized, compact student models that distill knowledge from a larger teacher model: the uedin-cuni submission to the wmt 2020 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Vikrant Goyal, Anoop Kunchukuttan, Rahul Kejriwal, Siddharth Jain, and Amit Bhagwat. 2020. Contact relatedness can help improve multilingual nmt: Microsoft stci-mt @ wmt20. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

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*, Virtual.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Jeremy Gwinnup and Tim Anderson. 2020. The afrl wmt20 news-translation systems. In *Proceedings of the Fifth 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 e-prints*, page arXiv:2001.09907.

Sami Ul Haq, Sadaf Abdul Rauf, Arsalan Shaukat, and Abdullah Saeed. 2020. Document level nmt of low-resource languages with backtranslation. In *Proceedings of WMT*.

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

François Hernandez and Vincent Nguyen. 2020. The ubiqus english-inuktiut system for wmt20. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, and Hajime Tsukada. 2010. Automatic Evaluation of Translation Quality for Distant Language Pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952, Cambridge, MA. Association for Computational Linguistics.

Eric Joanis, Rebecca Knowles, Roland Kuhn, Samuel Larkin, Patrick Littell, Chi-kiu Lo, Darlene Stewart, and Jeffrey Micher. 2020. The Nunavut Hansard Inuktitut–English parallel corpus 3.0 with preliminary machine translation results. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2562–2572, Marseille, France. European Language Resources Association.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Jiwan Kim, Soyoon Park, Sangha Kim, and Yoonjung Choi. 2020. An iterative knowledge transfer nmt system for wmt20 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Shun Kiyono, Takumi Ito, Ryuto Konno, Makoto Morishita, and Jun Suzuki. 2020. Tohoku-aip-ntt at wmt 2020 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Rebecca Knowles, Darlene Stewart, Samuel Larkin, and Patrick Littell. 2020. Nrc systems for the 2020 inuktitut-english news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Tom Kocmi. 2020. Cuni submission for inuktitut language in wmt news 2020. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Tom Kocmi, Tomasz Limisiewicz, and Gabriel Stanovsky. 2020. Gender coreference and bias evaluation at wmt 2020. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the wmt 2020 shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Christopher J. Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th*

Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

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.

Rihards Krišlauks and Mārcis Pinnis. 2020. Tilde at wmt 2020: News task systems. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Mateusz Krubiński, Marcin Chochowski, Bartłomiej Boczek, Mikołaj Koszowski, Adam Dobrowolski, Marcin Szymański, and Paweł Przybysz. 2020. Samsung r&d institute poland submission to wmt20 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Amit Kumar, Rupjyoti Baruah, Rajesh Kumar Mundotiya, and Anil Kumar Singh. 2020. Transformer-based neural machine translation system for hindi - marathi. In *Proceedings of WMT*.

Anoop Kunchukuttan. 2020a. Indowordnet parallel corpus.

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

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, and Sivaji Bandyopadhyay. 2020. Hindi-marathi cross lingual model. In *Proceedings of WMT*.

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. 2018. Has Neural Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In *EMNLP 2018*, Brussels, Belgium. Association for Computational Linguistics.

Zuchao Li, Hai Zhao, Rui Wang, Kehai Chen, Masao Utiyama, and Eiichiro Sumita. 2020. Sjtu-nict's supervised and unsupervised neural machine translation systems for the wmt20 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Zihan Liu, Genta Indra Winata, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2019. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems.

Nikola Ljubešić and Tomaž Erjavec. 2011. hrwac and slwac: Compiling web corpora for croatian and slovene. In *International Conference on Text, Speech and Dialogue*, pages 395–402. Springer.

Nikola Ljubešić and Filip Klubička. 2014. {bs, hr, sr} wac-web corpora of bosnian, croatian and serbian. In *Proceedings of the 9th web as corpus workshop (WaC-9)*, pages 29–35.

Qingsong Ma, Yvette Graham, Timothy Baldwin, and Qun Liu. 2017. Further investigation into reference bias in monolingual evaluation of machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2466–2475, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Lovish Madaan, Soumya Sharma, and Parag Singla. 2020. Transfer learning for related languages: Iit delhi's submissions to the wmt20 similar language translation task. In *Proceedings of WMT*.

Benjamin Marie, Raphael Rubino, and Atsushi Fujita. 2020. Combination of neural machine translation systems at wmt20. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Luıs Marujo, Nuno Grazina, Tiago Luıs, Wang Ling, Luısa Coheur, and Isabel Trancoso. 2011. BP2EP–Adaptation of Brazilian Portuguese texts to European Portuguese. In *Proceedings of EAMT*.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the wmt20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Fandong Meng, Jianhao Yan, Yijin Liu, Yuan Gao, Xianfeng Zeng, Qinsong Zeng, Peng Li, Ming Chen, Jie Zhou, Sifan Liu, and Hao Zhou. 2020. Wechat neural machine translation systems for wmt20. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Luis A. Menéndez-Salazar, Grigori Sidorov, and Marta R. Costa-Jussà. 2020. The ipn-cic team system submission for the wmt 2020 simmilar language task. In *Proceedings of WMT*.

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.

Alexander Molchanov. 2020. Promt systems for wmt 2020 shared news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, Online. 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.

Vandan Mujadia and Dipti Sharmaa. 2020. Nmt based similar language translation for hindi - marathi. In *Proceedings of WMT*.

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

Atul Kr. Ojha, Priya Rani, Akanksha Bansal, Bharathi Raja Chakravarthi, Ritesh Kumar, and John P. McCrae. 2020. Nuig-panlingua-kmi hindi–marathi mt systems for similar language translation task @ wmt 2020. In *Proceedings of WMT*.

Csaba Oravecz, Katina Bontcheva, László Tihanyi, David Kolovratnik, Bhavani Bhaskar, Adrien Lardilleux, Szymon Klocek, and Andreas Eisele. 2020. etranslation's submissions to the wmt 2020 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, Online. 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*, Brussels, Belgium. Association for Computational Linguistics.

Santanu Pal and Marcos Zampieri. 2020. Neural machine translation for similar languages: The case of indo-aryan languages. In *Proceedings of WMT*.

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, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Venkatesh Parthasarathy, Akshai Ramesh, Rejwanul Haque, and Andy Way. 2020. The adapt system description for the wmt20 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Ellie Pavlick, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. The language demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, 2(Feb):79–92.

Martin Popel. 2018. CUNI transformer neural MT system for WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 482–487, Belgium, Brussels. Association for Computational Linguistics.

Martin Popel. 2020. Cuni english-czech and english-polish systems in wmt20: Robust document-level training. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

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 Popovic. 2019. On reducing translation shifts in translations intended for MT evaluation. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 80–87, Dublin, Ireland. European Association for Machine Translation.

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)*.

Maja Popović and Alberto Poncelas. 2020. Neural machine translation between similar south-slavic languages. In *Proceedings of WMT*.

R. Pryzant, Y. Chung, D. Jurafsky, and D. Britz. 2017. JESC: Japanese-English Subtitle Corpus. *arXiv preprint arXiv:1710.10639*.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2019. The MuCoW Test Suite at WMT 2019: Automatically Harvested Multilingual Contrastive Word Sense Disambiguation Test Sets for Machine

Translation. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Kamalkumar Rathinasamy, Amanpreet Singh, Balaguru Sivasambagupta, Prajna Prasad Neerchal, and Vani Sivasankaran. 2020. Infosys machine translation system for wmt20 similar language translation task. In *Proceedings of WMT*.

Christian Roest, Lukas Edman, Gosse Minnema, Kevin Kelly, Jennifer Spenader, and Antonio Toral. 2020. Machine translation for english–inuktitut with segmentation,data acquisition and pre-training. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Yves Scherrer, Stig-Arne Grönroos, and Sami Virpioja. 2020a. The university of helsinki and aalto university submissions to the wmt 2020 news and low-resource translation tasks. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Yves Scherrer, Alessandro Raganato, and Jörg Tiedemann. 2020b. The mucow word sense disambiguation test suite at wmt 2020. In *Proceedings of the Fifth Conference on Machine Translation*, Online. 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.

Tingxun Shi, Shiyu Zhao, Xiaopu Li, Xiaoxue Wang, Qian Zhang, Di Ai, Dawei Dang, Xue Zhengshan, and JIE HAO. 2020. Oppo's machine translation systems for wmt20. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

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 Ralph Weischedel. 2006. A study of translation error rate with targeted human annotation. In *In Proceedings of the Association for Machine Transaltion in the Americas (AMTA 2006*.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020a. Findings of the wmt 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Lucia Specia, Zhenhao Li, Juan Pino, Vishrav Chaudhary, Francisco Guzmán, Graham Neubig, Nadir Durrani, Yonatan Belinkov, Philipp Koehn, Hassan Sajjad, Paul Michel, and Xian Li. 2020b. Findings of the wmt 2020 shared task on machine translation robustness. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

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

Jörg Tiedemann. 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*.

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

Dániel Varga, Péter Halaácsy, András Kornai, Voktor Nagy, László Németh, and Viktor Trón. 2005. Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005 Conference*, pages 590–596.

Tereza Vojtěchová, Michal Novák, Miloš Klouček, and Ondřej Bojar. 2019. SAO WMT19 Test Suite: Machine Translation of Audit Reports. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Daimeng Wei, Hengchao Shang, Zhanglin Wu, Zhengzhe Yu, Liangyou Li, Jiaxin Guo, Minghan Wang, Hao Yang, Lizhi Lei, Ying Qin, and Shiliang Sun. 2020a. Hw-tsc's participation in the wmt 2020 news translation shared task. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Xiangpeng Wei, Ping Guo, Yunpeng Li, Xingsheng Zhang, Luxi Xing, and Yue Hu. 2020b. Iie's neural machine translation systems for wmt20. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Liwei Wu, Xiao Pan, Zehui Lin, Yaoming ZHU, Mingxuan Wang, and Lei Li. 2020a. The volctrans machine translation system for wmt20. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Shuangzhi Wu, Xing Wang, Longyue Wang, Fangxu Liu, Jun Xie, Zhaopeng Tu, Shuming Shi, and Mu Li. 2020b. Tencent neural machine translation systems for the wmt20 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

ariel Xv. 2020. Russian-english bidirectional machine translation system. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Saumitra Yadav and Manish Shrivastava. 2020. A3-108 machine translation system for similar language translation shared task 2020. In *Proceedings of WMT*.

Lei Yu, Laurent Sartran, Po-Sen Huang, Wojciech Stokowiec, Domenic Donato, Srivatsan Srinivasan, Alek Andreev, Wang Ling, Sona Mokra, Agustin Dal Lago, Yotam Doron, Susannah Young, Phil Blunsom, and Chris Dyer. 2020. The deepmind chinese–english document translation system at wmt2020. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Jiacheng Zhang, Yanzhuo Ding, Shiqi Shen, Yong Cheng, Maosong Sun, Huan-Bo Luan, and Yang Liu. 2017. THUMT: an open source toolkit for neural machine translation. *CoRR*, abs/1706.06415.

Xiaoheng Zhang. 1998. Dialect MT: A Case Study Between Cantonese and Mandarin. In *Proceedings of ACL*.

Yuhao Zhang, Ziyang Wang, Runzhe Cao, Binghao Wei, Weiqiao Shan, Shuhan Zhou, Abudurexiti Reheman, Tao Zhou, Xin Zeng, Laohu Wang, Yongyu Mu, Jingnan Zhang, Xiaoqian Liu, Xuanjun Zhou, Yinqiao Li, Bei Li, Tong Xiao, and Jingbo Zhu. 2020. The niutrans machine translation systems for wmt20. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Vilém Zouhar, Tereza Vojtěchová, and Ondřej Bojar. 2020. Wmt20 document-level markable error exploration. In *Proceedings of the Fifth Conference on Machine Translation*, Online. Association for Computational Linguistics.

A Differences in Human Scores

Tables 33–50 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.

	CUNI-DocTransformer	OPPO	Online-B	CUNI-Transformer	Online-A	SRPOL	UEDIN-CUNI	CUNI-T2T-2018	Online-G	Online-Z	PROMT-NMT	zlabs-nlp
CUNI-DocTransformer	-	0.05	0.08	0.08⋆	0.10	0.15†	0.17‡	0.18‡	0.19‡	0.20‡	0.21‡	0.26‡
OPPO	-0.05	-	0.03	0.04	0.05	0.11⋆	0.12‡	0.14‡	0.14‡	0.15‡	0.17‡	0.21‡
Online-B	-0.08	-0.03	-	0.01	0.02	0.08	0.09†	0.11†	0.11†	0.12†	0.14†	0.18‡
CUNI-Transformer	-0.08	-0.04	-0.01	-	0.02	0.07	0.08⋆	0.10⋆	0.10†	0.11†	0.13†	0.18‡
Online-A	-0.10	-0.05	-0.02	-0.02	-	0.05	0.07⋆	0.08⋆	0.09⋆	0.10†	0.11†	0.16‡
SRPOL	-0.15	-0.11	-0.08	-0.07	-0.05	-	0.01	0.03	0.03	0.04	0.06	0.10†
UEDIN-CUNI	-0.17	-0.12	-0.09	-0.08	-0.07	-0.01	-	0.02	0.02	0.03	0.05	0.09⋆
CUNI-T2T-2018	-0.18	-0.14	-0.11	-0.10	-0.08	-0.03	-0.02	-	0.00	0.02	0.03	0.08
Online-G	-0.19	-0.14	-0.11	-0.10	-0.09	-0.03	-0.02	0.00	-	0.01	0.03	0.07
Online-Z	-0.20	-0.15	-0.12	-0.11	-0.10	-0.04	-0.03	-0.02	-0.01	-	0.01	0.06
PROMT-NMT	-0.21	-0.17	-0.14	-0.13	-0.11	-0.06	-0.05	-0.03	-0.03	-0.01	-	0.05
zlabs-nlp	-0.26	-0.21	-0.18	-0.18	-0.16	-0.10	-0.09	-0.08	-0.07	-0.06	-0.05	-
score	0.12	0.07	0.04	0.03	0.02	-0.04	-0.05	-0.07	-0.07	-0.08	-0.09	-0.14
rank	1–12	1–12	1–12	1–12	1–12	1–12	1–12	1–12	1–12	1–12	1–12	1–12

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

	VolcTrans	DiDi-NLP	WeChat-AI	Tencent-Translation	Online-B	DeepMind	OPPO	THUNLP	SJTU-NICT	Huawei-TSC	Online-A	HUMAN	Online-G	dong-NMT	zlabs-NLP	Online-Z	WMTBiomedBaseline
VolcTrans	-	0.01⋆	0.02†	0.04†	0.04⋆	0.05†	0.05‡	0.07†	0.09‡	0.10‡	0.12‡	0.13‡	0.17‡	0.18‡	0.21‡	0.24‡	0.43‡
DiDi-NLP	-0.01	-	0.01	0.03	0.03	0.04	0.04⋆	0.06	0.07	0.09‡	0.11†	0.12‡	0.16‡	0.17‡	0.19‡	0.22‡	0.42‡
WeChat-AI	-0.02	-0.01	-	0.01	0.02	0.03	0.03	0.05	0.06	0.08†	0.09⋆	0.11‡	0.15‡	0.16‡	0.18‡	0.21‡	0.41‡
Tencent-Translation	-0.04	-0.03	-0.01	-	0.00	0.01	0.01	0.04	0.05	0.06⋆	0.08	0.09†	0.13‡	0.14‡	0.17‡	0.20‡	0.40‡
Online-B	-0.04	-0.03	-0.02	0.00	-	0.01	0.01	0.03	0.04	0.06†	0.08⋆	0.09‡	0.13‡	0.14‡	0.17‡	0.20‡	0.39‡
DeepMind	-0.05	-0.04	-0.03	-0.01	-0.01	-	0.00	0.02	0.03	0.05†	0.07⋆	0.08†	0.12‡	0.13‡	0.16‡	0.19‡	0.38‡
OPPO	-0.05	-0.04	-0.03	-0.01	-0.01	0.00	-	0.02	0.03	0.05⋆	0.07	0.08†	0.12‡	0.13‡	0.16‡	0.19‡	0.38‡
THUNLP	-0.07	-0.06	-0.05	-0.04	-0.03	-0.02	-0.02	-	0.01	0.03†	0.04	0.06†	0.10‡	0.11‡	0.13‡	0.16‡	0.36‡
SJTU-NICT	-0.09	-0.07	-0.06	-0.05	-0.04	-0.03	-0.03	-0.01	-	0.02⋆	0.03	0.05†	0.09‡	0.09‡	0.12‡	0.15‡	0.35‡
Huawei-TSC	-0.10	-0.09	-0.08	-0.06	-0.06	-0.05	-0.05	-0.03	-0.02	-	0.02	0.03	0.07⋆	0.08†	0.11†	0.13‡	0.33‡
Online-A	-0.12	-0.11	-0.09	-0.08	-0.08	-0.07	-0.07	-0.04	-0.03	-0.02	-	0.01	0.05†	0.06‡	0.09‡	0.12‡	0.32‡
HUMAN	-0.13	-0.12	-0.11	-0.09	-0.09	-0.08	-0.08	-0.06	-0.05	-0.03	-0.01	-	0.04	0.05⋆	0.08†	0.11‡	0.30‡
Online-G	-0.17	-0.16	-0.15	-0.13	-0.13	-0.12	-0.12	-0.10	-0.09	-0.07	-0.05	-0.04	-	0.01	0.03	0.06⋆	0.26‡
dong-NMT	-0.18	-0.17	-0.16	-0.14	-0.14	-0.13	-0.13	-0.11	-0.09	-0.08	-0.06	-0.05	-0.01	-	0.03	0.06	0.25‡
zlabs-NLP	-0.21	-0.19	-0.18	-0.17	-0.17	-0.16	-0.16	-0.13	-0.12	-0.11	-0.09	-0.08	-0.03	-0.03	-	0.03	0.23‡
Online-Z	-0.24	-0.22	-0.21	-0.20	-0.20	-0.19	-0.19	-0.16	-0.15	-0.13	-0.12	-0.11	-0.06	-0.06	-0.03	-	0.20‡
WMTBiomedBaseline	-0.43	-0.42	-0.41	-0.40	-0.39	-0.38	-0.38	-0.36	-0.35	-0.33	-0.32	-0.30	-0.26	-0.25	-0.23	-0.20	-
score	0.10	0.09	0.08	0.06	0.06	0.05	0.05	0.03	0.02	0.00	-0.02	-0.03	-0.07	-0.08	-0.11	-0.14	-0.33
rank	1	2–16	2–16	2–16	2–16	2–16	2–16	2–16	2–16	2–16	2–16	2–16	2–16	2–16	2–16	2–16	17

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

	VolcTrans	OPPO	HUMAN	Tohoku-AIP-NTT	Online-A	Online-G	PROMT-NMT	Online-B	UEDIN	Online-Z	WMTBiomedBaseline	zlabs-NLP	yolo
VolcTrans	-	0.01	0.04⋆	0.05	0.05	0.06	0.06⋆	0.06⋆	0.10⋆	0.14‡	0.31‡	0.33‡	1.85‡
OPPO	-0.01	-	0.03⋆	0.04	0.04	0.05	0.05⋆	0.05⋆	0.09	0.13‡	0.30‡	0.33‡	1.84‡
HUMAN	-0.04	-0.03	-	0.01	0.01	0.01	0.02	0.02	0.06	0.10	0.27‡	0.29‡	1.80‡
Tohoku-AIP-NTT	-0.05	-0.04	-0.01	-	0.00	0.01	0.01	0.01	0.05	0.09†	0.26‡	0.28‡	1.80‡
Online-A	-0.05	-0.04	-0.01	0.00	-	0.01	0.01	0.01	0.05	0.09†	0.26‡	0.28‡	1.80‡
Online-G	-0.06	-0.05	-0.01	-0.01	-0.01	-	0.00	0.01	0.04	0.09†	0.25‡	0.28‡	1.79‡
PROMT-NMT	-0.06	-0.05	-0.02	-0.01	-0.01	0.00	-	0.00	0.04	0.09	0.25‡	0.28‡	1.79‡
Online-B	-0.06	-0.05	-0.02	-0.01	-0.01	-0.01	0.00	-	0.04	0.08	0.25‡	0.27‡	1.78‡
UEDIN	-0.10	-0.09	-0.06	-0.05	-0.05	-0.04	-0.04	-0.04	-	0.05⋆	0.21‡	0.24‡	1.75‡
Online-Z	-0.14	-0.13	-0.10	-0.09	-0.09	-0.09	-0.09	-0.08	-0.05	-	0.16‡	0.19‡	1.70‡
WMTBiomedBaseline	-0.31	-0.30	-0.27	-0.26	-0.26	-0.25	-0.25	-0.25	-0.21	-0.16	-	0.03	1.54‡
zlabs-NLP	-0.33	-0.33	-0.29	-0.28	-0.28	-0.28	-0.28	-0.27	-0.24	-0.19	-0.03	-	1.51‡
yolo	-1.85	-1.84	-1.80	-1.80	-1.80	-1.79	-1.79	-1.78	-1.75	-1.70	-1.54	-1.51	-
score	0.23	0.22	0.19	0.18	0.18	0.17	0.17	0.17	0.13	0.09	-0.08	-0.11	-1.62
rank	1–9	1–9	1–9	1–9	1–9	1–9	1–9	1–9	1–9	10	11–12	11–12	13

Table 35: Head to head comparison for German→English systems

	ONLINE-G	ONLINE-A	OPPO	eTRANSLATION	PROMT-NMT	ONLINE-B	HUMAN	ARIEL XV	AFRL	DIDI-NLP	ONLINE-Z	ZLABS-NLP
ONLINE-G	-	0.01	0.01	0.02	0.03	0.05†	0.06	0.08⋆	0.10†	0.14‡	0.15‡	0.28‡
ONLINE-A	-0.01	-	0.00	0.01	0.02	0.04	0.05	0.07	0.09⋆	0.13†	0.14†	0.27‡
OPPO	-0.01	0.00	-	0.01	0.02	0.04⋆	0.05	0.07⋆	0.09†	0.13†	0.13‡	0.27‡
eTRANSLATION	-0.02	-0.01	-0.01	-	0.01	0.03⋆	0.04	0.06⋆	0.08⋆	0.12†	0.13†	0.26‡
PROMT-NMT	-0.03	-0.02	-0.02	-0.01	-	0.02	0.03	0.05	0.07	0.11⋆	0.12⋆	0.25‡
ONLINE-B	-0.05	-0.04	-0.04	-0.03	-0.02	-	0.01	0.03	0.05	0.09	0.09	0.23‡
HUMAN	-0.06	-0.05	-0.05	-0.04	-0.03	-0.01	-	0.02	0.04	0.08⋆	0.08⋆	0.22‡
ARIEL XV	-0.08	-0.07	-0.07	-0.06	-0.05	-0.03	-0.02	-	0.02	0.06	0.06	0.20‡
AFRL	-0.10	-0.09	-0.09	-0.08	-0.07	-0.05	-0.04	-0.02	-	0.04	0.05	0.18‡
DIDI-NLP	-0.14	-0.13	-0.13	-0.12	-0.11	-0.09	-0.08	-0.06	-0.04	-	0.01	0.14‡
ONLINE-Z	-0.15	-0.14	-0.13	-0.13	-0.12	-0.09	-0.08	-0.06	-0.05	-0.01	-	0.13‡
ZLABS-NLP	-0.28	-0.27	-0.27	-0.26	-0.25	-0.23	-0.22	-0.20	-0.18	-0.14	-0.13	-
score	0.12	0.11	0.11	0.10	0.10	0.07	0.06	0.04	0.03	-0.02	-0.02	-0.15
rank	1–11	1–11	1–11	1–11	1–11	1–11	1–11	1–11	1–11	1–11	1–11	12

Table 36: Head to head comparison for Russian→English systems

	TOHOKU-AIP-NTT	NIUTRANS	OPPO	NICT-KYOTO	ONLINE-B	ONLINE-A	eTRANSLATION	ZLABS-NLP	ONLINE-G	ONLINE-Z
TOHOKU-AIP-NTT	-	0.04	0.10	0.10⋆	0.12‡	0.16‡	0.16‡	0.39‡	0.40‡	0.42‡
NIUTRANS	-0.04	-	0.06	0.06⋆	0.08‡	0.12‡	0.13‡	0.35‡	0.37‡	0.39‡
OPPO	-0.10	-0.06	-	0.00	0.02⋆	0.06⋆	0.07⋆	0.30‡	0.31‡	0.33‡
NICT-KYOTO	-0.10	-0.06	0.00	-	0.02	0.06	0.06	0.29‡	0.30‡	0.32‡
ONLINE-B	-0.12	-0.08	-0.02	-0.02	-	0.04	0.05	0.28‡	0.29‡	0.31‡
ONLINE-A	-0.16	-0.12	-0.06	-0.06	-0.04	-	0.01	0.23‡	0.25‡	0.27‡
eTRANSLATION	-0.16	-0.13	-0.07	-0.06	-0.05	-0.01	-	0.23‡	0.24‡	0.26‡
ZLABS-NLP	-0.39	-0.35	-0.30	-0.29	-0.28	-0.23	-0.23	-	0.01	0.03
ONLINE-G	-0.40	-0.37	-0.31	-0.30	-0.29	-0.25	-0.24	-0.01	-	0.02
ONLINE-Z	-0.42	-0.39	-0.33	-0.32	-0.31	-0.27	-0.26	-0.03	-0.02	-
score	0.18	0.15	0.09	0.08	0.07	0.03	0.02	-0.21	-0.22	-0.24
rank	1–7	1–7	1–7	1–7	1–7	1–7	1–7	8–10	8–10	8–10

Table 37: Head to head comparison for Japanese→English systems

	SRPOL	ONLINE-G	NICT-RUI	ONLINE-B	SJTU-NICT	ONLINE-A	OPPO	ONLINE-Z	CUNI-TRANSFORMER	NICT-KYOTO	VOLCTRANS	PROMT-NMT	TILDE	ZLABS-NLP
SRPOL	-	0.03	0.04	0.04⋆	0.05	0.05⋆	0.08†	0.12†	0.13‡	0.17‡	0.17‡	0.18‡	0.20‡	0.26‡
ONLINE-G	-0.03	-	0.00	0.00	0.01	0.01	0.05	0.09⋆	0.10⋆	0.13†	0.14‡	0.14‡	0.17‡	0.23‡
NICT-RUI	-0.04	0.00	-	0.00	0.01	0.01	0.05⋆	0.09†	0.10†	0.13‡	0.14‡	0.14‡	0.17‡	0.23‡
ONLINE-B	-0.04	0.00	0.00	-	0.01	0.01	0.04	0.09	0.10	0.13⋆	0.14†	0.14†	0.17‡	0.22‡
SJTU-NICT	-0.05	-0.01	-0.01	-0.01	-	0.00	0.04	0.08⋆	0.09†	0.12†	0.13‡	0.13‡	0.16‡	0.22‡
ONLINE-A	-0.05	-0.01	-0.01	-0.01	0.00	-	0.03	0.08	0.09⋆	0.12⋆	0.12†	0.13†	0.15‡	0.21‡
OPPO	-0.08	-0.05	-0.05	-0.04	-0.04	-0.03	-	0.04	0.05	0.09	0.09⋆	0.10⋆	0.12†	0.18‡
ONLINE-Z	-0.12	-0.09	-0.09	-0.09	-0.08	-0.08	-0.04	-	0.01	0.04	0.05	0.05	0.08†	0.14†
CUNI-TRANSFORMER	-0.13	-0.10	-0.10	-0.10	-0.09	-0.09	-0.05	-0.01	-	0.03	0.04	0.04	0.07⋆	0.13†
NICT-KYOTO	-0.17	-0.13	-0.13	-0.13	-0.12	-0.12	-0.09	-0.04	-0.03	-	0.00	0.01	0.03	0.09⋆
VOLCTRANS	-0.17	-0.14	-0.14	-0.14	-0.13	-0.12	-0.09	-0.05	-0.04	0.00	-	0.01	0.03	0.09
PROMT-NMT	-0.18	-0.14	-0.14	-0.14	-0.13	-0.13	-0.10	-0.05	-0.04	-0.01	-0.01	-	0.02	0.08
TILDE	-0.20	-0.17	-0.17	-0.17	-0.16	-0.15	-0.12	-0.08	-0.07	-0.03	-0.03	-0.02	-	0.06
ZLABS-NLP	-0.26	-0.23	-0.23	-0.22	-0.22	-0.21	-0.18	-0.14	-0.13	-0.09	-0.09	-0.08	-0.06	-
score	0.13	0.10	0.10	0.09	0.09	0.08	0.05	0.01	-0.00	-0.04	-0.04	-0.05	-0.07	-0.13
rank	1–14	1–14	1–14	1–14	1–14	1–14	1–14	1–14	1–14	1–14	1–14	1–14	1–14	1–14

Table 38: Head to head comparison for Polish→English systems

	GTCOM	OPPO	ONLINE-B	FACEBOOK-AI	NIUTRANS	VOLCTRANS	ONLINE-Z	ZLABS-NLP	MICROSOFT-STC-INDIA	UEDIN	ONLINE-A	DCU	ONLINE-G	TALP-UPC
GTCOM	-	0.00	0.03	0.03	0.05	0.09	0.20‡	0.20‡	0.22‡	0.22‡	0.27‡	0.28‡	0.60‡	0.65‡
OPPO	0.00	-	0.03	0.03	0.05	0.09	0.20‡	0.20‡	0.22‡	0.22‡	0.27‡	0.28‡	0.60‡	0.65‡
ONLINE-B	-0.03	-0.03	-	0.00	0.03	0.06	0.17‡	0.17‡	0.19‡	0.20†	0.24‡	0.25‡	0.57‡	0.63‡
FACEBOOK-AI	-0.03	-0.03	0.00	-	0.02	0.06⋆	0.17‡	0.17‡	0.19‡	0.19‡	0.24‡	0.25‡	0.57‡	0.62‡
NIUTRANS	-0.05	-0.05	-0.03	-0.02	-	0.03	0.14‡	0.15†	0.17‡	0.17‡	0.22‡	0.23‡	0.55‡	0.60‡
VOLCTRANS	-0.09	-0.09	-0.06	-0.06	-0.03	-	0.11†	0.12†	0.13‡	0.14†	0.18‡	0.19‡	0.51‡	0.57‡
ONLINE-Z	-0.20	-0.20	-0.17	-0.17	-0.14	-0.11	-	0.01	0.02	0.03	0.07	0.08	0.41‡	0.46‡
ZLABS-NLP	-0.20	-0.20	-0.17	-0.17	-0.15	-0.12	-0.01	-	0.02	0.02	0.07	0.08	0.40‡	0.45‡
MICROSOFT-STC-INDIA	-0.22	-0.22	-0.19	-0.19	-0.17	-0.13	-0.02	-0.02	-	0.00	0.05	0.06	0.38‡	0.43‡
UEDIN	-0.22	-0.22	-0.20	-0.19	-0.17	-0.14	-0.03	-0.02	0.00	-	0.05	0.06	0.38‡	0.43‡
ONLINE-A	-0.27	-0.27	-0.24	-0.24	-0.22	-0.18	-0.07	-0.07	-0.05	-0.05	-	0.01	0.33‡	0.38‡
DCU	-0.28	-0.28	-0.25	-0.25	-0.23	-0.19	-0.08	-0.08	-0.06	-0.06	-0.01	-	0.32‡	0.37‡
ONLINE-G	-0.60	-0.60	-0.57	-0.57	-0.55	-0.51	-0.41	-0.40	-0.38	-0.38	-0.33	-0.32	-	0.05⋆
TALP-UPC	-0.65	-0.65	-0.63	-0.62	-0.60	-0.57	-0.46	-0.45	-0.43	-0.43	-0.38	-0.37	-0.05	-
score	0.20	0.20	0.18	0.17	0.15	0.12	0.01	0.00	-0.02	-0.02	-0.07	-0.08	-0.40	-0.45
rank	1–6	1–6	1–6	1–6	1–6	1–6	7–12	7–12	7–12	7–12	7–12	7–12	13	14

Table 39: Head to head comparison for Tamil→English systems

	NiuTrans	Facebook-AI	CUNI-Transfer	Groningen	SRPOL	Helsinki	NRC	UEDIN	UQAM-TanLe	NICT-Kyoto	OPPO
NiuTrans	-	0.00	0.07⋆	0.07⋆	0.10†	0.10⋆	0.11†	0.11‡	0.12†	0.16‡	0.20‡
Facebook-AI	0.00	-	0.07⋆	0.07⋆	0.10⋆	0.10⋆	0.11†	0.11‡	0.12†	0.16‡	0.20‡
CUNI-Transfer	-0.07	-0.07	-	0.00	0.03	0.03	0.04	0.05	0.05	0.09⋆	0.13‡
Groningen	-0.07	-0.07	0.00	-	0.02	0.03	0.04	0.04	0.05	0.09⋆	0.13†
SRPOL	-0.10	-0.10	-0.03	-0.02	-	0.01	0.02	0.02	0.02	0.07⋆	0.11†
Helsinki	-0.10	-0.10	-0.03	-0.03	-0.01	-	0.01	0.01	0.02	0.06⋆	0.10†
NRC	-0.11	-0.11	-0.04	-0.04	-0.02	-0.01	-	0.00	0.01	0.05	0.09⋆
UEDIN	-0.11	-0.11	-0.05	-0.04	-0.02	-0.01	0.00	-	0.01	0.05	0.09⋆
UQAM-TanLe	-0.12	-0.12	-0.05	-0.05	-0.02	-0.02	-0.01	-0.01	-	0.04	0.08⋆
NICT-Kyoto	-0.16	-0.16	-0.09	-0.09	-0.07	-0.06	-0.05	-0.05	-0.04	-	0.04
OPPO	-0.20	-0.20	-0.13	-0.13	-0.11	-0.10	-0.09	-0.09	-0.08	-0.04	-
score	0.17	0.17	0.10	0.10	0.07	0.07	0.06	0.05	0.05	0.01	-0.04
rank	1–2	1–2	3–11	3–11	3–11	3–11	3–11	3–11	3–11	3–11	3–11

Table 40: Head to head comparison for Inuktitut→English systems

	Online-B	GTCOM	Huawei-TSC	VolcTrans	OPPO	Online-Z
Online-B	-	0.01	0.05⋆	0.14‡	0.20‡	0.23‡
GTCOM	-0.01	-	0.04	0.13‡	0.19‡	0.22‡
Huawei-TSC	-0.05	-0.04	-	0.09⋆	0.15‡	0.18‡
VolcTrans	-0.14	-0.13	-0.09	-	0.06⋆	0.09†
OPPO	-0.20	-0.19	-0.15	-0.06	-	0.03
Online-Z	-0.23	-0.22	-0.18	-0.09	-0.03	-
score	0.03	0.02	-0.02	-0.11	-0.16	-0.20
rank	1–3	1–3	1–3	4	5–6	5–6

Table 41: Head to head comparison for Pashto→English systems

	Online-B	GTCOM	Huawei-TSC	VolcTrans	OPPO	Online-Z	Online-G
Online-B	-	0.02	0.03	0.22‡	0.38‡	0.39‡	0.45‡
GTCOM	-0.02	-	0.01	0.19‡	0.36‡	0.37‡	0.43‡
Huawei-TSC	-0.03	-0.01	-	0.18‡	0.35‡	0.36‡	0.42‡
VolcTrans	-0.22	-0.19	-0.18	-	0.16‡	0.18‡	0.23‡
OPPO	-0.38	-0.36	-0.35	-0.16	-	0.01	0.07
Online-Z	-0.39	-0.37	-0.36	-0.18	-0.01	-	0.06
Online-G	-0.45	-0.43	-0.42	-0.23	-0.07	-0.06	-
score	0.17	0.15	0.14	-0.05	-0.21	-0.22	-0.28
rank	1–3	1–3	1–3	4	5–7	5–7	5–7

Table 42: Head to head comparison for Khmer→English systems

	HUMAN-B	HUMAN-A	OPPO	TENCENT-TRANSLATION	HUAWEI-TSC	NIUTRANS	SJTU-NICT	VOLCTRANS	ONLINE-B	ONLINE-A	DONG-NMT	ONLINE-Z	ONLINE-G	ZLABS-NLP
HUMAN-B	-	0.04‡	0.12‡	0.15‡	0.15‡	0.16‡	0.18‡	0.19‡	0.29‡	0.33‡	0.43‡	0.43‡	0.45‡	0.49‡
HUMAN-A	-0.04	-	0.08‡	0.11‡	0.11‡	0.13‡	0.14‡	0.16‡	0.25‡	0.29‡	0.39‡	0.39‡	0.41‡	0.45‡
OPPO	-0.12	-0.08	-	0.03‡	0.03‡	0.04‡	0.06‡	0.07‡	0.16‡	0.21‡	0.31‡	0.31‡	0.32‡	0.36‡
TENCENT-TRANSLATION	-0.15	-0.11	-0.03	-	0.01	0.02	0.03⋆	0.05	0.14‡	0.18‡	0.28‡	0.29‡	0.30‡	0.34‡
HUAWEI-TSC	-0.15	-0.11	-0.03	-0.01	-	0.01	0.03⋆	0.04	0.13‡	0.17‡	0.28‡	0.28‡	0.29‡	0.33‡
NIUTRANS	-0.16	-0.13	-0.04	-0.02	-0.01	-	0.02	0.03	0.12‡	0.16‡	0.27‡	0.27‡	0.28‡	0.32‡
SJTU-NICT	-0.18	-0.14	-0.06	-0.03	-0.03	-0.02	-	0.01	0.10‡	0.15‡	0.25‡	0.25‡	0.27‡	0.30‡
VOLCTRANS	-0.19	-0.16	-0.07	-0.05	-0.04	-0.03	-0.01	-	0.09‡	0.13‡	0.24‡	0.24‡	0.25‡	0.29‡
ONLINE-B	-0.29	-0.25	-0.16	-0.14	-0.13	-0.12	-0.10	-0.09	-	0.04‡	0.15‡	0.15‡	0.16‡	0.20‡
ONLINE-A	-0.33	-0.29	-0.21	-0.18	-0.17	-0.16	-0.15	-0.13	-0.04	-	0.11‡	0.11‡	0.12‡	0.16‡
DONG-NMT	-0.43	-0.39	-0.31	-0.28	-0.28	-0.27	-0.25	-0.24	-0.15	-0.11	-	0.00	0.01	0.05†
ONLINE-Z	-0.43	-0.39	-0.31	-0.29	-0.28	-0.27	-0.25	-0.24	-0.15	-0.11	0.00	-	0.01	0.05⋆
ONLINE-G	-0.45	-0.41	-0.32	-0.30	-0.29	-0.28	-0.27	-0.25	-0.16	-0.12	-0.01	-0.01	-	0.04⋆
ZLABS-NLP	-0.49	-0.45	-0.36	-0.34	-0.33	-0.32	-0.30	-0.29	-0.20	-0.16	-0.05	-0.05	-0.04	-
score	0.57	0.53	0.45	0.42	0.41	0.40	0.39	0.37	0.28	0.24	0.14	0.14	0.12	0.08
rank	1	2	3	4–8	4–8	4–8	4–8	4–8	9	10	11–13	11–13	11–13	14

Table 43: Head to head comparison for English→Chinese systems

	HUMAN	CUNI-DOCTRANSFORMER	OPPO	SRPOL	CUNI-T2T-2018	ETRANSLATION	CUNI-TRANSFORMER	UEDIN-CUNI	ONLINE-B	ONLINE-Z	ONLINE-A	ONLINE-G	ZLABS-NLP
HUMAN	-	0.11‡	0.12‡	0.15‡	0.20‡	0.21‡	0.22‡	0.33‡	0.61‡	0.64‡	0.65‡	0.87‡	1.41‡
CUNI-DOCTRANSFORMER	-0.11	-	0.01	0.04†	0.09‡	0.11‡	0.11‡	0.22‡	0.50‡	0.53‡	0.54‡	0.76‡	1.31‡
OPPO	-0.12	-0.01	-	0.03⋆	0.08‡	0.10‡	0.10‡	0.22‡	0.49‡	0.52‡	0.53‡	0.75‡	1.30‡
SRPOL	-0.15	-0.04	-0.03	-	0.05⋆	0.06‡	0.07†	0.18‡	0.46‡	0.49‡	0.50‡	0.72‡	1.26‡
CUNI-T2T-2018	-0.20	-0.09	-0.08	-0.05	-	0.02	0.02	0.14‡	0.41‡	0.44‡	0.45‡	0.67‡	1.22‡
ETRANSLATION	-0.21	-0.11	-0.10	-0.06	-0.02	-	0.01	0.12‡	0.39‡	0.42‡	0.43‡	0.66 ‡	1.20‡
CUNI-TRANSFORMER	-0.22	-0.11	-0.10	-0.07	-0.02	-0.01	-	0.11‡	0.39‡	0.42‡	0.43‡	0.65‡	1.19‡
UEDIN-CUNI	-0.33	-0.22	-0.22	-0.18	-0.14	-0.12	-0.11	-	0.27‡	0.30‡	0.31‡	0.54‡	1.08‡
ONLINE-B	-0.61	-0.50	-0.49	-0.46	-0.41	-0.39	-0.39	-0.27	-	0.03	0.04	0.26‡	0.81‡
ONLINE-Z	-0.64	-0.53	-0.52	-0.49	-0.44	-0.42	-0.42	-0.30	-0.03	-	0.01	0.23‡	0.78‡
ONLINE-A	-0.65	-0.54	-0.53	-0.50	-0.45	-0.43	-0.43	-0.31	-0.04	-0.01	-	0.22‡	0.77‡
ONLINE-G	-0.87	-0.76	-0.75	-0.72	-0.67	-0.66	-0.65	-0.54	-0.26	-0.23	-0.22	-	0.54‡
ZLABS-NLP	-1.41	-1.31	-1.30	-1.26	-1.22	-1.20	-1.19	-1.08	-0.81	-0.78	-0.77	-0.54	-
score	0.65	0.55	0.54	0.51	0.46	0.44	0.43	0.32	0.05	0.02	0.01	-0.22	-0.76
rank	1	2–3	2–3	4	5–7	5–7	5–7	8	9–11	9–11	9–11	12	13

Table 44: Head to head comparison for English→Czech systems

	HUMAN-B	OPPO	TOHOKU-AIP-NTT	HUMAN-A	ONLINE-B	TENCENT-TRANSLATION	VOLCTRANS	ONLINE-A	ETRANSLATION	HUMAN-C	AFRL	UEDIN	PROMT-NMT	ONLINE-Z	ONLINE-G	ZLABS-NLP	WMTBIOMEDBASELINE
HUMAN-B	-	0.07⋆	0.10‡	0.12⋆	0.15‡	0.18‡	0.24‡	0.25‡	0.26‡	0.27‡	0.31‡	0.32‡	0.32‡	0.44‡	0.69‡	0.85‡	0.91‡
OPPO	-0.07	-	0.03	0.05	0.08⋆	0.11‡	0.17‡	0.17‡	0.18‡	0.20‡	0.24‡	0.24‡	0.25‡	0.37‡	0.61‡	0.77‡	0.83‡
TOHOKU-AIP-NTT	-0.10	-0.03	-	0.02	0.05	0.08⋆	0.14‡	0.15‡	0.16‡	0.17‡	0.21‡	0.22‡	0.22‡	0.34‡	0.59‡	0.75‡	0.81‡
HUMAN-A	-0.12	-0.05	-0.02	-	0.03⋆	0.06†	0.12‡	0.12‡	0.13‡	0.15‡	0.19‡	0.19‡	0.20‡	0.32‡	0.57‡	0.72‡	0.78‡
ONLINE-B	-0.15	-0.08	-0.05	-0.03	-	0.03	0.09‡	0.09‡	0.10‡	0.12‡	0.16‡	0.17‡	0.17‡	0.29‡	0.54‡	0.69‡	0.75‡
TENCENT-TRANSLATION	-0.18	-0.11	-0.08	-0.06	-0.03	-	0.06‡	0.06‡	0.07†	0.09‡	0.12‡	0.13‡	0.14‡	0.26‡	0.50‡	0.66‡	0.72‡
VOLCTRANS	-0.24	-0.17	-0.14	-0.12	-0.09	-0.06	-	0.00	0.01	0.03	0.07	0.08†	0.08†	0.20‡	0.45‡	0.60‡	0.66‡
ONLINE-A	-0.25	-0.17	-0.15	-0.12	-0.09	-0.06	0.00	-	0.01	0.02	0.06	0.07†	0.08†	0.20‡	0.44‡	0.60‡	0.66‡
ETRANSLATION	-0.26	-0.18	-0.16	-0.13	-0.10	-0.07	-0.01	-0.01	-	0.01	0.05	0.06†	0.06‡	0.19‡	0.43‡	0.59‡	0.65‡
HUMAN-C	-0.27	-0.20	-0.17	-0.15	-0.12	-0.09	-0.03	-0.02	-0.01	-	0.04	0.05⋆	0.05⋆	0.17‡	0.42‡	0.58‡	0.64‡
AFRL	-0.31	-0.24	-0.21	-0.19	-0.16	-0.12	-0.07	-0.06	-0.05	-0.04	-	0.01	0.01⋆	0.13‡	0.38‡	0.54‡	0.60‡
UEDIN	-0.32	-0.24	-0.22	-0.19	-0.17	-0.13	-0.08	-0.07	-0.06	-0.05	-0.01	-	0.00	0.13†	0.37‡	0.53‡	0.59‡
PROMT-NMT	-0.32	-0.25	-0.22	-0.20	-0.17	-0.14	-0.08	-0.08	-0.06	-0.05	-0.01	0.00	-	0.12‡	0.37‡	0.53‡	0.59‡
ONLINE-Z	-0.44	-0.37	-0.34	-0.32	-0.29	-0.26	-0.20	-0.20	-0.19	-0.17	-0.13	-0.13	-0.12	-	0.25‡	0.40‡	0.46‡
ONLINE-G	-0.69	-0.61	-0.59	-0.57	-0.54	-0.50	-0.45	-0.44	-0.43	-0.42	-0.38	-0.37	-0.37	-0.25	-	0.16	0.22†
ZLABS-NLP	-0.85	-0.77	-0.75	-0.72	-0.69	-0.66	-0.60	-0.60	-0.59	-0.58	-0.54	-0.53	-0.53	-0.40	-0.16	-	0.06
WMTBIOMEDBASELINE	-0.91	-0.83	-0.81	-0.78	-0.75	-0.72	-0.66	-0.66	-0.65	-0.64	-0.60	-0.59	-0.59	-0.46	-0.22	-0.06	-
score	0.57	0.49	0.47	0.45	0.42	0.39	0.33	0.32	0.31	0.30	0.26	0.25	0.25	0.13	-0.12	-0.28	-0.34
rank	1	2–6	2–6	2–6	2–6	2–6	7–13	7–13	7–13	7–13	7–13	7–13	7–13	14	15–17	15–17	15–17

Table 45: Head to head comparison for English→German systems

	HUMAN	MULTILINGUAL-UBIQUS	CUNI-TRANSFER	NRC	FACEBOOK-AI	NICT_KYOTO	GRONINGEN	HELSINKI	SRPOL	UQAM-TANLE	UEDIN	OPPO
HUMAN	-	0.15	0.17†	0.21‡	0.21‡	0.21†	0.24‡	0.28‡	0.29‡	0.49‡	0.49‡	0.96‡
MULTILINGUAL-UBIQUS	-0.15	-	0.02⋆	0.06‡	0.06‡	0.06⋆	0.09‡	0.13‡	0.14‡	0.34‡	0.34‡	0.81‡
CUNI-TRANSFER	-0.17	-0.02	-	0.04†	0.04	0.04	0.07†	0.11‡	0.13‡	0.32‡	0.33‡	0.79‡
NRC	-0.21	-0.06	-0.04	-	0.00	0.01	0.03	0.07	0.09	0.29‡	0.29‡	0.75‡
FACEBOOK-AI	-0.21	-0.06	-0.04	0.00	-	0.00	0.03	0.07†	0.09†	0.28‡	0.29‡	0.75‡
NICT-KYOTO	-0.21	-0.06	-0.04	-0.01†	0.00	-	0.02†	0.07‡	0.08‡	0.28‡	0.28‡	0.75‡
GRONINGEN	-0.24	-0.09	-0.07	-0.03	-0.03	-0.02	-	0.04	0.06	0.26‡	0.26‡	0.72‡
HELSINKI	-0.28	-0.13	-0.11	-0.07	-0.07	-0.07	-0.04	-	0.01	0.21‡	0.21‡	0.68‡
SRPOL	-0.29	-0.14	-0.13	-0.09	-0.09	-0.08	-0.06	-0.01	-	0.20‡	0.20‡	0.67‡
UQAM-TANLE	-0.49	-0.34	-0.32	-0.29	-0.28	-0.28	-0.26	-0.21	-0.20	-	0.00	0.47‡
UEDIN	-0.49	-0.34	-0.33	-0.29	-0.29	-0.28	-0.26	-0.21	-0.20	0.00	-	0.47‡
OPPO	-0.96	-0.81	-0.79	-0.75	-0.75	-0.75	-0.72	-0.68	-0.67	-0.47	-0.47	-
score	0.57	0.42	0.41	0.37	0.37	0.36	0.34	0.30	0.28	0.08	0.08	-0.38
rank	1–2	1–2	3–9	3–9	3–9	3–9	3–9	3–9	3–9	10–11	10–11	12

Table 46: Head to head comparison for English→Inuktitut systems

	HUMAN	NIUTRANS	TOHOKU-AIP-NTT	OPPO	ENMT	NICT-KYOTO	ONLINE-A	ONLINE-B	ZLABS-NLP	ONLINE-Z	SJTU-NICT	ONLINE-G
HUMAN	-	0.07†	0.08†	0.08†	0.08†	0.20‡	0.23‡	0.24‡	0.42‡	0.54‡	0.71‡	0.74‡
NIUTRANS	-0.07	-	0.01	0.01	0.01	0.13‡	0.15‡	0.17‡	0.34‡	0.47‡	0.63‡	0.67‡
TOHOKU-AIP-NTT	-0.08	-0.01	-	0.00	0.00	0.12‡	0.15‡	0.16‡	0.34‡	0.46‡	0.63‡	0.66‡
OPPO	-0.08	-0.01	0.00	-	0.00	0.12‡	0.15‡	0.16‡	0.34‡	0.46‡	0.63‡	0.66‡
ENMT	-0.08	-0.01	0.00	0.00	-	0.12‡	0.14‡	0.16‡	0.33‡	0.46‡	0.62‡	0.66‡
NICT-KYOTO	-0.20	-0.13	-0.12	-0.12	-0.12	-	0.03	0.04⋆	0.22‡	0.34‡	0.51‡	0.54‡
ONLINE-A	-0.23	-0.15	-0.15	-0.15	-0.14	-0.03	-	0.01	0.19‡	0.32‡	0.48‡	0.51‡
ONLINE-B	-0.24	-0.17	-0.16	-0.16	-0.16	-0.04	-0.01	-	0.18‡	0.30‡	0.47‡	0.50‡
ZLABS-NLP	-0.42	-0.34	-0.34	-0.34	-0.33	-0.22	-0.19	-0.18	-	0.13‡	0.29‡	0.32‡
ONLINE-Z	-0.54	-0.47	-0.46	-0.46	-0.46	-0.34	-0.32	-0.30	-0.13	-	0.16‡	0.20‡
SJTU-NICT	-0.71	-0.63	-0.63	-0.63	-0.62	-0.51	-0.48	-0.47	-0.29	-0.16	-	0.03
ONLINE-G	-0.74	-0.67	-0.66	-0.66	-0.66	-0.54	-0.51	-0.50	-0.32	-0.20	-0.03	-
score	0.58	0.50	0.50	0.50	0.49	0.38	0.35	0.34	0.16	0.03	-0.13	-0.16
rank	1	2–5	2–5	2–5	2–5	6–8	6–8	6–8	9	10	11–12	11–12

Table 47: Head to head comparison for English→Japanese systems

	HUMAN	SRPOL	ETRANSLATION	VOLCTRANS	TILDE	ONLINE-G	OPPO	NICT-KYOTO	TILDE	CUNI-TRANSFORMER	ONLINE-B	SJTU-NICT	ONLINE-A	ONLINE-Z	ZLABS-NLP
HUMAN	-	0.18‡	0.24‡	0.29‡	0.32‡	0.36‡	0.36‡	0.37‡	0.40‡	0.42‡	0.44‡	0.45‡	0.58‡	0.73‡	1.21‡
SRPOL	-0.18	-	0.06⋆	0.11†	0.14‡	0.18‡	0.18‡	0.19‡	0.22‡	0.24‡	0.26‡	0.27‡	0.40‡	0.55‡	1.03‡
ETRANSLATION	-0.24	-0.06	-	0.05	0.09†	0.12†	0.12‡	0.14†	0.16‡	0.18‡	0.20‡	0.22‡	0.34‡	0.49‡	0.97‡
VOLCTRANS	-0.29	-0.11	-0.05	-	0.03	0.07	0.07‡	0.08	0.11‡	0.13†	0.15‡	0.16‡	0.29‡	0.44‡	0.92‡
TILDE	-0.32	-0.14	-0.09	-0.03	-	0.03	0.04⋆	0.05	0.08⋆	0.09	0.11‡	0.13‡	0.25‡	0.41‡	0.89‡
ONLINE-G	-0.36	-0.18	-0.12	-0.07	-0.03	-	0.01	0.02	0.04	0.06	0.08†	0.10†	0.22‡	0.38‡	0.85‡
OPPO	-0.36	-0.18	-0.12	-0.07	-0.04	-0.01	-	0.01	0.04	0.06	0.07⋆	0.09⋆	0.21‡	0.37‡	0.85‡
NICT-KYOTO	-0.37	-0.19	-0.14	-0.08	-0.05	-0.02	-0.01⋆	-	0.03⋆	0.04	0.06‡	0.08‡	0.20‡	0.36‡	0.84‡
TILDE	-0.40	-0.22	-0.16	-0.11	-0.08	-0.04	-0.04	-0.03	-	0.02	0.04	0.05	0.18‡	0.33‡	0.81‡
CUNI-TRANSFORMER	-0.42	-0.24	-0.18	-0.13	-0.09	-0.06	-0.06	-0.04	-0.02	-	0.02⋆	0.04⋆	0.16‡	0.31‡	0.79‡
ONLINE-B	-0.44	-0.26	-0.20	-0.15	-0.11	-0.08	-0.07	-0.06	-0.04	-0.02	-	0.02	0.14‡	0.30‡	0.77‡
SJTU-NICT	-0.45	-0.27	-0.22	-0.16	-0.13	-0.10	-0.09	-0.08	-0.05	-0.04	-0.02	-	0.12†	0.28‡	0.76‡
ONLINE-A	-0.58	-0.40	-0.34	-0.29	-0.25	-0.22	-0.21	-0.20	-0.18	-0.16	-0.14	-0.12	-	0.16‡	0.63‡
ONLINE-Z	-0.73	-0.55	-0.49	-0.44	-0.41	-0.38	-0.37	-0.36	-0.33	-0.31	-0.30	-0.28	-0.16	-	0.48‡
ZLABS-NLP	-1.21	-1.03	-0.97	-0.92	-0.89	-0.85	-0.85	-0.84	-0.81	-0.79	-0.77	-0.76	-0.63	-0.48	-
score	0.67	0.49	0.43	0.38	0.35	0.32	0.31	0.30	0.27	0.26	0.24	0.22	0.10	-0.06	-0.54
rank	1	2	3–8	3–8	3–8	3–8	3–8	3–8	9–10	9–10	11–12	11–12	13	14	15

Table 48: Head to head comparison for English→Polish systems

	HUMAN	ONLINE-G	OPPO	ARIEL XV	ONLINE-B	PROMT-NMT	DIDI-NLP	ONLINE-A	ZLABS-NLP	ONLINE-Z
HUMAN	-	0.21‡	0.22‡	0.28‡	0.35‡	0.43‡	0.46‡	0.60‡	0.65‡	0.67‡
ONLINE-G	-0.21	-	0.01	0.06⋆	0.13‡	0.22‡	0.25‡	0.39‡	0.43‡	0.46‡
OPPO	-0.22	-0.01	-	0.06	0.13‡	0.21‡	0.24‡	0.38‡	0.43‡	0.45‡
ARIEL XV	-0.28	-0.06	-0.06	-	0.07†	0.15‡	0.18‡	0.32‡	0.37‡	0.39‡
ONLINE-B	-0.35	-0.13	-0.13	-0.07	-	0.08†	0.11‡	0.25‡	0.30‡	0.32‡
PROMT-NMT	-0.43	-0.22	-0.21	-0.15	-0.08	-	0.03⋆	0.17‡	0.22‡	0.24‡
DIDI-NLP	-0.46	-0.25	-0.24	-0.18	-0.11	-0.03	-	0.14‡	0.19‡	0.21‡
ONLINE-A	-0.60	-0.39	-0.38	-0.32	-0.25	-0.17	-0.14	-	0.05	0.07†
ZLABS-NLP	-0.65	-0.43	-0.43	-0.37	-0.30	-0.22	-0.19	-0.05	-	0.02
ONLINE-Z	-0.67	-0.46	-0.45	-0.39	-0.32	-0.24	-0.21	-0.07	-0.02	-
score	0.68	0.47	0.46	0.40	0.34	0.25	0.22	0.08	0.04	0.01
rank	1	2–4	2–4	2–4	5	6	7	8–10	8–10	8–10

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

	HUMAN	FACEBOOK-AI	GTCOM	ONLINE-B	OPPO	ONLINE-A	VOLCTRANS	ONLINE-Z	ZLABS-NLP	MICROSOFT-STC-INDIA	UEDIN	GRONINGEN	DCU	TALP-UPC	ONLINE-G	SJTU-NICT
HUMAN	-	0.10‡	0.25‡	0.27‡	0.28‡	0.31‡	0.34‡	0.44‡	0.45‡	0.47‡	0.53‡	0.61‡	0.77‡	1.17‡	1.48‡	1.58‡
FACEBOOK-AI	-0.10	-	0.15‡	0.17‡	0.18‡	0.21‡	0.24‡	0.34‡	0.36‡	0.37‡	0.43‡	0.51‡	0.67‡	1.07‡	1.38‡	1.48‡
GTCOM	-0.25	-0.15	-	0.02	0.03	0.06	0.09	0.19‡	0.21‡	0.22‡	0.28‡	0.36‡	0.52‡	0.92‡	1.23‡	1.33‡
ONLINE-B	-0.27	-0.17	-0.02	-	0.01	0.03	0.07	0.17‡	0.18‡	0.19‡	0.26‡	0.34‡	0.50‡	0.90‡	1.21‡	1.31‡
OPPO	-0.28	-0.18	-0.03	-0.01	-	0.02	0.06	0.15‡	0.17‡	0.18‡	0.25‡	0.33‡	0.49‡	0.89‡	1.20‡	1.30‡
ONLINE-A	-0.31	-0.21	-0.06	-0.03	-0.02	-	0.03	0.13‡	0.15†	0.16‡	0.23‡	0.30‡	0.46‡	0.86‡	1.17‡	1.28‡
VOLCTRANS	-0.34	-0.24	-0.09	-0.07	-0.06	-0.03	-	0.10‡	0.12⋆	0.13†	0.19‡	0.27‡	0.43‡	0.83‡	1.14‡	1.24‡
ONLINE-Z	-0.44	-0.34	-0.19	-0.17	-0.15	-0.13	-0.10	-	0.02	0.03	0.09	0.17⋆	0.33‡	0.73‡	1.04‡	1.15‡
ZLABS-NLP	-0.45	-0.36	-0.21	-0.18	-0.17	-0.15	-0.12	-0.02	-	0.01	0.08	0.15†	0.31‡	0.71‡	1.02‡	1.13‡
MICROSOFT-STC-INDIA	-0.47	-0.37	-0.22	-0.19	-0.18	-0.16	-0.13	-0.03	-0.01	-	0.07	0.14⋆	0.30‡	0.70‡	1.01‡	1.12‡
UEDIN	-0.53	-0.43	-0.28	-0.26	-0.25	-0.23	-0.19	-0.09	-0.08	-0.07	-	0.08	0.24‡	0.64‡	0.95‡	1.05‡
GRONINGEN	-0.61	-0.51	-0.36	-0.34	-0.33	-0.30	-0.27	-0.17	-0.15	-0.14	-0.08	-	0.16‡	0.56‡	0.87‡	0.97‡
DCU	-0.77	-0.67	-0.52	-0.50	-0.49	-0.46	-0.43	-0.33	-0.31	-0.30	-0.24	-0.16	-	0.40‡	0.71‡	0.81‡
TALP-UPC	-1.17	-1.07	-0.92	-0.90	-0.89	-0.86	-0.83	-0.73	-0.71	-0.70	-0.64	-0.56	-0.40	-	0.31‡	0.41‡
ONLINE-G	-1.48	-1.38	-1.23	-1.21	-1.20	-1.17	-1.14	-1.04	-1.02	-1.01	-0.95	-0.87	-0.71	-0.31	-	0.10⋆
SJTU-NICT	-1.58	-1.48	-1.33	-1.31	-1.30	-1.28	-1.24	-1.15	-1.13	-1.12	-1.05	-0.97	-0.81	-0.41	-0.10	-
score	0.76	0.66	0.51	0.49	0.48	0.46	0.42	0.33	0.31	0.30	0.23	0.15	-0.01	-0.41	-0.72	-0.82
rank	1	2	3–7	3–7	3–7	3–7	3–7	8–12	8–12	8–12	8–12	8–12	13	14	15	16

Table 50: Head to head comparison for English→Tamil 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 Translator Brief: Paragraph-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 paragraph or newline boundaries and blank lines.** The source texts are formatted as short paragraphs separated by blank lines. We need this formatting preserved so that we can align the sources and translations.
- 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). We need the translations to be returned in the same format, ideally with utf8 encoding. If you prefer to receive the text in a different format, then please let us know as we may be able to accommodate it.