Findings of the 2019 Conference on Machine Translation (WMT19)

Loïc Barrault
Le Mans Université

Ondřej Bojar
Charles University

Marta R. Costa-jussà
UPC

Christian Federmann
Microsoft Cloud + AI

Mark Fishel
University of Tartu

Yvette Graham
Dublin City University

Barry Haddow
University of Edinburgh

Matthias Huck
LMU Munich

Philipp Koehn
JHU / University of Edinburgh

Shervin Malmasi
Harvard Medical School

Christof Monz
University of Amsterdam

Mathias Müller
University of Zurich

Santanu Pal
Saarland University

Matt Post
JHU

Marcos Zampieri
University of Wolverhampton

Abstract

This paper presents the results of the premier shared task organized alongside the Conference on Machine Translation (WMT) 2019. Participants were asked to build machine translation systems for any of 18 language pairs, to be evaluated on a test set of news stories. The main metric for this task is human judgment of translation quality. The task was also opened up to additional test suites to probe specific aspects of translation.

1 Introduction

The Fourth Conference on Machine Translation (WMT) held at ACL 2019[1] hosts a number of shared tasks on various aspects of machine translation. This conference builds on 13 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).

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:

- biomedical translation (Bawden et al., 2019b)
- automatic post-editing (Chatterjee et al., 2019)
- metrics (Ma et al., 2019)
- quality estimation (Fonseca et al., 2019)
- parallel corpus filtering (Koehn et al., 2019)
- robustness (Li et al., 2019b)

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 held 18 translation tasks this year, between English and each of Chinese, Czech (into Czech only), German, Finnish, Lithuanian, and Russian. New this year were Gujarati↔English and Kazakh↔English. Both pose a lesser resourced data condition on challenging language pairs. System outputs for each task were evaluated both automatically and manually.

This year the news translation task had two additional sub-tracks: an unsupervised language pair (German→Czech) and a language pair not involving English (German↔French). Both sub-tracks were included into the general list of news translation submissions and are described in more detail in the corresponding subsections of Section 2.

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. This year, the official manual evaluation metric is again based on judgments of adequacy on a 100-point scale, a method 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-

[1] http://www.statmt.org/wmt19/

[2] http://statmt.org/wmt19/results.html

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers (Day 1) pages 1–61
Florence, Italy, August 1-2, 2019.

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 focussing on some particular aspect of translation quality. A brief overview of the test suites is provided in Section 4.

2 News Translation Task

The recurring WMT task examines translation between English and other languages in the news domain. As in the previous year, we include Chinese, Czech, German, Finnish and Russian (into and out of English, except for Czech were only out of English was included). New language pairs for this year were Gujarati, Lithuanian and Kazakh (to and from English), and French-German. We also used German-Czech (joining the corresponding parts of the English-X test sets) for the unsupervised subtask.

2.1 Test Data

The test data for this year's task (except for the French-German set) was selected from online news sources, as in previous years, with translation produced specifically for the task. For language pairs that had appeared before at WMT (and so had previous years' data for development testing) we selected approximately 2000 sentences in each of the languages in the pair and translated them into the other language. The source English sentences were common across all test sets. For the new language pairs (i.e. English-Gujarati, English-Kazakh and English-Lithuanian) we released development sets at the start of the campaign, consisting of approximately 1000 sentences in each language in the pair, translated into the other language. For Gujarati-English the development set was selected from online news in the same way as the test set, whereas for Kazakh-English the development set was selected (and removed) from the news-commentary training set. The test sets for these new language pairs was half the size of the test sets of the existing language pairs.

Different to previous years, all test sets (except for French-German and German-Czech) only included naturally occurring text on the source side. In previous years, the way we produced an English-X test set was to take 1500 sentences of English text, translate these into language X, then take 1500 sentences in language X, and translated them into English. These 3000 translation pairs were then used for the English-X task, and for the X-English task, meaning that 50% of the sentences in each test has "translationese" on the source side, potentially leading to distortions in automatic and human evaluation (Graham et al., 2019a). This year, we did not include such "flipped" test data in the test sets, meaning that the English-X and X-English sets were non-overlapping.

The composition of the test documents is shown in Table 1, the size of the test sets in terms of sentence pairs and words is given in Figure 2.

The translation of the test sets was sponsored by the EU H2020 projects Bergamot and GoURMET (English-Czech and Gujarati-English respectively), by Yandex (Kazakh-English and Russian-English), Microsoft (Chinese-English and German-English), Tilde (Lithuanian-English), the University of Helsinki (Finnish-English) and Lingua Custodia[3] (a part of French-German test set).

The translations into Czech were carried out by the agency Překlady textu, s.r.o.[4] with the instructions for translators as given to all agencies:

- preserve line and document boundaries,
- translate from scratch, without post-editing,
- translate as literally as possible, but ensure that the translation is still a fluent sentence in the target language,
- do not add or remove information from the translations, and do not add translator's comments.
- The point is to have a linguistically nice document, but to be matching the original text as closely as possible in terms of segmentation into sentences (e.g. we don't want 3 English sentences combined into 1 long Czech complex sentence).

[3] http://www.linguacustodia.finance/
[4] http://www.preklady-textu.cz/

Europarl Parallel Corpus

	Czech ↔ English		Finnish ↔ English		German ↔ English		Lithuanian↔ English		French ↔ German	
Sentences	645,241		1,835,071		1,825,741		631,309		1,726,419	
Words	14,948,882	17,380,337	35,766,351	50,233,589	48,125,049	50,506,042	13,448,546	17,070,302	46,014,903	41,000,331
Distinct words	172,450	63,287	677,673	112,751	371,743	113,958	237,740	62,885	388,613	616,702

News Commentary Parallel Corpus

	Czech ↔ English		German ↔ English		Russian ↔ English	
Sentences	240,243		329,506		281,184	
Words	5,372,690	5,938,908	8,363,213	8,295,418	7,132,754	7,447,684
Distinct words	172,215	68,966	197,056	80,623	194,808	76,953

	Chinese ↔ English		Kazakh ↔ English		French ↔ German	
Sentences	311,922		7,475		256,226	
Words	–	7,926,131	157,171	193,101	8,049,218	6,607,025
Distinct words	–	75,955	24,676	13,982	82,740	171,410

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		Lithuanian ↔ English	
Sentences	31,358,551		5,862,521		1,368,691	
Words	559,348,288	598,362,329	89,066,831	93,943,773	20,992,360	23,111,861
Distinct Words	8,081,990	4,888,665	1,477,399	1,108,068	723,940	495,311

	Finnish ↔ English		Russian ↔ English		French ↔ German	
Sentences	3,944,929		12,061,155		7,222,574	
Words	55,245,472	66,352,625	182,325,667	210,770,856	145,190,707	123,205,701
Distinct Words	1,787,403	944,140	2,958,831	2,385,075	1,534,068	2,368,682

EU Press Release Parallel Corpus

	German ↔ English		Finnish ↔ English		Lithuanian↔ English	
Sentences	1,480,789		583,223		213,173	
Words	29,458,773	30,097,541	8,052,607	11,244,602	4,097,713	4,817,655
Distinct words	399,545	165,084	315,394	94,979	106,603	53,239

Chinese Parallel Corpora

	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

Yandex 1M Parallel Corpus

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

CzEng v1.7 Parallel Corpus

	Czech ↔ English	
Sentences	57,065,358	
Words	667,091,440	751,312,654
Distinct	2,592,850	1,639,658

WikiTitles Parallel Corpus

	Czech ↔ English		German ↔ English		Finnish ↔ English		Gujarati↔ English	
Sentences	362,014		1,305,135		376,572		11,670	
Words	862,719	924,948	2,817,660	3,271,223	761,213	912,044	23,780	24,098
Distinct	197,743	168,449	618,723	525,023	232,236	183,285	11,557	10,400

	Kazakh↔ English		Lithuanian↔ English		Russian↔ English		Chinese↔ English	
Sentences	117,041		132,182		1,032,343		765,674	
Words	189,565	231,166	286,837	304,043	2,786,728	2,793,609	–	2,031,512
Distinct	94,525	86,587	95,004	83,404	481,018	410,112	–	341,166

United Nations Parallel Corpus

	Russian ↔ English		Chinese ↔ English	
Sentences	23,239,280		15,886,041	
Words	482,966,738	524,719,646	–	372,612,596
Distinct	3,857,656	2,737,469	–	1,981,413

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 provided tokenizer and IndicNLP (https://github.com/anoopkunchukuttan/indic_nlp_library) for Gujarati.

Crawled Kazakh Parallel Data

	Kazakh ↔ English		Russian ↔ English	
Sentences	97,654		5,063,666	
Words	1,224,971	1,524,384	111,492,772	115,950,305
Distinct	89,500	39,704	1,022,853	774,991

Crawled Gujarati-English Parallel Data

	The Bible		Localisation		Indian Govt.		Wikipedia	
Sentences	7,807		107,637		10,650		18,033	
Words	228,113	206,440	763,521	750,659	154,364	177,141	370,972	373,491
Distinct	15,623	5,945	15,406	8,549	23,489	16,361	57,431	32,227

Monolingual Wikipedia Data

	Gujarati	Kazakh	Lithuanian
Sentences	384,485	2,179,180	2,059,198
Words	6,779,645	28,130,741	31,006,475
Distinct words	373,840	1,115,320	970,696

News Language Model Data

	English	German	Czech	Russian	Finnish
Sentences 199,900,557		275,690,481	72,157,988	80,148,714	16,834,066
Words	4,611,843,099	4,922,055,629	1,193,459,840	1,461,279,309	213,048,421
Distinct words	6,910,887	34,747,433	4,668,868	4,771,311	5,084,937

	Gujarati	Kazakh	Lithuanian	French	Chinese
Sentences	244,919	772,892	375,206	76,848,192	1,749,968
Words	3,776,100	13,172,313	6,782,918	1,858,333,964	–
Distinct words	183,425	506,923	288,266	3,376,105	–

Document-Split News LM Data (not dedudped)

	English	German	Czech
Sentences	419,796,579	533,619,919	92,388,432
Words	9,305,189,308	9,520,383,021	1,512,084,445
Distinct words	6,813,799	34,668,232	4,582,601

Common Crawl Language Model Data

	English	German	Czech	Russian	Finnish
Sent.	3,074,921,453	2,872,785,485	333,498,145	1,168,529,851	157,264,161
Words	65,128,419,540	65,154,042,103	6,694,811,063	23,313,060,950	2,935,402,545
Dist.	342,760,462	339,983,035	50,162,437	101,436,673	47,083,545

	Chinese	Lithuanian	Kazakh	Gujarati	French
Sent.	1,672,324,647	103,103,449	10,862,371	3,729,406	
Words	–	2,907,519,260	261,518,626	80,120,267	
Dist.	–	25,343,195	4,381,617	2,068,064	

Test Sets

	Chinese → EN		EN → Chinese		EN → Czech		Finnish → EN		EN → Finnish		German → EN	
Sentences.	2000		1997		1997		1996		1997		2000	
Words	–	80,666	48,021	–	48,021	43,860	24,797	36,809	48,021	38,068	36,141	39,561
Distinct words	–	7,939	7,372	–	7,372	11,537	10,454	5,763	7,372	12,789	8,763	6,764

	EN → German		Gujarati → EN		EN → Gujarati		Kazakh → EN		EN → Kazakh		Lithuanian→ EN	
Sentences.	1997		1016		998		1000		998		1000	
Words	48,021	49,069	15,691	17,950	24,074	22,285	16,259	20,376	24,074	19,142	20,027	26,020
Distinct words	7,372	9,659	5,013	3,388	4,772	6,558	6,200	3,761	4,772	7,113	7,178	4,424

	EN→ Lithuanian		Russian → EN		EN → Russian		German → Czech		French ↔ German	
Sentences.	998		2000		1997		1997		1701	
Words	24,074	20,603	35,821	43,158	48,021	48,298	49,779	43,860	46,216	36,563
Distinct words	4,772	7,046	10,564	6,311	7,372	12,385	9,502	11,537	5,942	7,042

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

Language	Sources (Number of Documents)
Chinese	Chinanews (111), Macao Govt. (4), QQ (10), Reuters (31), RFI (2), Tsrus (5)
English I	ABC News (3), BBC (12), CBS News (2), CNBC (3), CNN (3), Daily Mail (9), Euronews (3), Guardian (3), Independent (3), News Week (6), NY Times (4), Reuters (3), Russia Today (1), The Scotsman (3), The Telegraph (2), UPI (2)
English II	ABC News (3), BBC (6), CBS News (4), CNBC (2), CNN (3), Daily Mail (2), Euronews (2), Fox News (1), Guardian (2), Independent (1), News Week (5), NY Times (4), Reuters (9), Russia Today (4), The Scotsman (6), The Telegraph (4), The Local (1), UPI (2)
Finnish	ESS (8), Helsinginsanomat (12), Iltalehti (33), Iltasanomat (34), Kaleva (19), Kansanuutiset (1), Karjalainen (26), Kotiseutu Uutiset (1)
German	Abdendzeitung München (9), Abendzeitung Nürnberg (1), Aachener Nachrichten (7), Augsburger Allgemine (2), Bergdorfer Zeitung (2), Braunschweiger Zeiting (2), Cuxhavener Nachrichten (1), Come On (2), Der Standart (9), Deutsche Welle (1), Duelmener Zeitung (7), Euronews (2), Frankfurter Neue Presse (2), Frankfurter Rundschau (4), Freipresse (1), Geinhaüser Tageblatt (1), Gmünder Tagespost (1), Göttinger Tageblatt (2), Handelsblatt (3), Hannoversche Allgemeine Zeitung (1), Hersfelder Zeitung (2), HNA (2), Infranken (5), In Süd Thüringen (3), Kieler Nachrichten (6), Merkur Online (5), Morgen Post (1), Nachrichten (4), N TV (3), NW News (1), NZZ (6), OE24 (5), PAZ Online (1), Passauer Neue Presse (1), Rhein Zeitung (1), Rheinische Poste (1), Salzburg (3), Schwarzwälder Bote (2), Söster Anzeiger (2), Südkurier (1), Usinger Anzeiger (1), Westfaelischer Anzeige (2), Welt (2), Wienerzeitung (2), Westfaelische Nachrichten (18), Zeit (1), Zeitungsverlag Waiblingen (2)
Gujarati	ABP Asmita (13), BBC (3), Divya Bhaskar (20), Global Gujarati News (13), Web Dunia (21)
Kazakh	7Kun (4), Aktobe Gazeti (3), Alkyn (4), Astana Akshamy (6), Atyray (1), Kazakh Adabieti (1), Egemen (5), Jaskazaq (11), Akorda/Kazinform (34), SN.kz (5), Zamedia (1)
Lithuanian	Delfi (22), Diena (25), Lietuvos Zinios (7), TV3 (12), Voruta (2), VZ (8)
Russian	AIF.ru (14), Altapress (4), Argumenti (3), Euronews (13), Fakty (9), Gazeta (7), Infox (3), Izvestiya (38), Kommersant (12), Lenta (14), Nezavisimaya Gazeta (8), Moskovskij Komsomolets (19), Parlamentskaya Gazeta (1), Rossiskaya Gazeta (1), ERR (1), Sovetskij Sport (31), Vedomosti (1), Nasha Versiya (1), Vesti (14), Za Rulyom (2)

Table 1: Composition of the test sets. English I was used for all language pairs, whereas English II was used for all except Gujarati, Kazakh and Lithuanian. 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.

2.2 Training Data

As in past years we provided parallel corpora to train translation models, monolingual corpora to train language models, and development sets to tune system parameters.

This year, we proposed document-level evaluation for the English-German and English-Czech tasks. We therefore attempted to provide training corpora with document boundaries intact wherever possible. We produced new versions of the Europarl corpora with document boundaries, an updated version of news-commentary with document boundaries, and a release of the Rapid corpus for German-English with document boundaries intact. The CzEng[5] already included context for each sentence, so we did not update it. We also produced a WikiTitle corpus this year for all language pairs, and allowed the use of a new ParaCrawl corpus (v3). The UN, CommonCrawl and Yandex corpora were unchanged since last year.

For Gujarati-English, we allowed several extra parallel corpora (the Bible, a localisation corpus from Opus, the Emille corpus, a Wikipedia corpus and a crawled corpus specifically for this task), as well as encouraging participants to experiment with the HindEnCorp[6] for transfer learning.

For Kazakh-English, we released a crawled corpus (from KazakhTV) prepared by Bagdat Myrzakhmetov of Nazarbayev University as well as a much larger Kazakh-Russian corpus for transfer learning or pivoting.

We released new monolingual news crawls for each of the languages used in the task. For German and Czech, we released versions of these with the document boundaries intact, for participants wishing to experiment with document-level models.

Some statistics about the training materials are given in Figures 1 and 2.

2.3 Unsupervised Sub-Task

Following up on the unsupervised learning challenge from last year, we again invited participants to build *unsupervised machine translation* systems without the use of any parallel training corpora.

While WMT has been (and is) providing considerable amounts of bitext for most of the language pairs covered in its shared tasks on machine translation of news, there is however still a shortage of available parallel resources between

[5] http://ufal.mff.cuni.cz/czeng/czeng17

[6] http://ufallab.ms.mff.cuni.cz/~bojar/hindencorp/

lots of combinations of two human languages. Bridging through a global hub language—such as English—can be a solution in scenarios where no bitext exists between two languages but parallel corpora with the hub language are at hand for each of the two. This "pivot translation" approach of cascading source–English and English–target MT is well-established. More recent research on unsupervised translation, on the other hand, seeks to altogether eliminate the need for parallel training data. Unsupervised translation techniques should be capable of learning translation correspondences from only monolingual data in two different languages, thus potentially offering a solution to machine translation between each and every possible pair of written human languages.

Previous year's evaluation had indicated that, unsurprisingly, unsupervised translation clearly lags behind supervised translation. But we had also seen promising early-stage research results which seemed to suggest that the difficult task of unsupervised learning in machine translation may not be impossible to solve in the long run. When acceptable quality can be reached with unsupervised methods, these methods will likely not directly compete with supervised translation, but rather be deployed to cover language pairs where supervised translation is inapplicable due to a lack of parallel data.

The language pair for the WMT19 unsupervised sub-task was German–Czech. Only the German→Czech translation direction was evaluated, not the Czech→German direction. German is a compounding language, and German and Czech are both morphologically rich. Linguistic peculiarities on both the source *and* the target side impose difficulties other than for last year's languages, where we paired Turkish, Estonian, and German each with English for the unsupervised sub-task. By choosing German–Czech, we hope to simulate practical application scenarios for fully unsupervised translation. However, note that there actually is German–Czech parallel data, e.g. from European parliamentary proceedings. German–English and English–Czech bitexts likewise exist in large amounts. We asked the participants to avoid any of these corpora, as well as any monolingual or parallel data for other languages and language pairs. Permissible training data for the unsupervised sub-task were only the monolingual corpora from the constrained monolingual WMT News Crawls of German and Czech. Last years' parallel dev and test sets (from the development tarball[7]) were allowed for bootstrapping purposes. Since they contain a few thousand sentences of high-quality German–Czech parallel text, we advised participants to make only very moderate use of this data. Using it directly as a training corpus was strongly discouraged, but we wanted to provide system builders with a means to evaluate and track progress internally during system development. We also did not prohibit its use for lightweight (hyper-)parameter optimization.

Seven German→Czech unsupervised machine translation systems were submitted and marked as primary submissions by the participating teams. The unsupervised system submissions were evaluated along with four online systems for the German→Czech language pairs, which we assume are all supervised MT engines. The official results of the human evaluation are presented in Table 12 (Section 3).

2.4 EUElections German→French and French→German Sub-Tasks

The second new sub-task this year included translating news data between French and German (both directions) on the topic of the European Elections. We collected a development and test set from online news websites. Articles were originally in French or in German. Statistics of the corpora a presented in the following table. In order to analyse the impact of the original source language of document on systems' performance, we computed the METEOR scores on the full corpus (FULL), on the sentences from articles initially written in French (second column) or in German (third column). Results are shown in the Tables 3 and 4. One can notice some differences depending on the language direction. While the performance of the systems when translating from French to German seems to heavily depend on the

	#lines	#token FR	#token DE
dev2019 FULL	1512	33833	28733
- source FR	462	11081	10890
- source DE	1050	22752	17843
test2019 FULL	1701	38154	31560
- source FR	335	7678	7195
- source DE	1366	30476	24365

Table 2: Statistics of the French↔German dev and test sets with breakdown statistics based on the source language.

[7] http://data.statmt.org/wmt19/translation-task/dev.tgz

Systems	FULL	source FR	source DE
MSRA.MADL	47.3	38.3	50.0
eTranslation	45.4	37.4	47.8
LIUM	43.7	37.5	45.5
MLLP-UPV	41.5	36.4	43.0
onlineA	40.8	35.4	42.3
TartuNLP	39.2	34.8	40.5
onlineB	39.1	35.3	40.2
onlineY	39.0	34.7	40.2
onlineG	38.5	34.6	39.7
onlineX	38.1	35.6	38.8

Table 3: French→German Meteor scores.

Systems	FULL	source FR	source DE
MSRA.MADL	52.0	51.9	52.0
LinguaCustodia	51.3	52.5	51.0
MLLP_UPV	49.5	49.9	49.4
Kyoto_University_T2T	48.8	49.7	48.6
LIUM	48.3	46.5	48.7
onlineY	47.5	43.7	48.4
onlineB	46.4	43.7	47.0
TartuNLP	46.3	45.0	46.7
onlineA	45.3	43.7	45.8
onlineX	42.7	41.6	42.9
onlineG	41.7	40.9	41.9

Table 4: German→French Meteor scores. Green cells highlight the systems performing equally when source text is in either language. The gray cells show that the TartuNLP system performs better with French source text relatively to its overall score.

original language of the document, this is less the case for the German to French direction. These results suggest that the German text produced by translating French documents is somewhat different from the German text originally produced even though native German translators were involved in the process. This is of course not new and is related to *translationese* (Koppel and Ordan, 2011). As shown in Table 2, only one fifth of the test corpus originates from French documents. With this in mind, Table 4 suggests that the *translationese* is less obvious for French text.

For next year, we plan to produce additional data with documents created during and after the elections.

2.5 Submitted Systems

In 2019, we received a total of 153 submissions. The participating institutions are listed in Table 5 and detailed in the rest of this section. Each system did not necessarily appear in all translation tasks. We also included online MT systems (originating from 5 services), which we anonymized as ONLINE-A,B,G,X,Y.

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 sub-section, we provide brief details of the submitted systems, for those in cases where the authors provided such details.

2.5.1 AFRL

AFRL-SYSCOMB19 **(Gwinnup et al., 2019)** is a system combination of a Marian ensemble system, two distinct OpenNMT systems, a Sockeye-based Elastic Weight Consolidation system, and one Moses phrase-based system.

AFRL-EWC **(Gwinnup et al., 2019)** is a Sockeye Transformer system trained with the default network configuration as described in Vaswani et al. (2017). The model is trained using the prepared parallel corpus used in other AFRL systems. A fine-tuning corpus is created from the 2014–2017 WMT Russian–English test sets. EWC is applied as described in Thompson et al. (2019). The final submission is an ensemble decode of the four best-performing checkpoints from a single training run when scoring newstest2018.

2.5.2 APERTIUM-FIN-ENG **(Pirinen, 2019)**

APERTIUM-FIN-ENG is a standard shallow rule-based machine translation using Apertium.

2.5.3 APPRENTICE-C **(Li and Specia, 2019)**

APPRENTICE-C is a RNN-based encoder-decoder with pre-trained embedding enhanced by character information. The system is trained on 10.38M Chinese-English sentence pairs after tokenization, filtering by alignment and BPE . Pre-trained embedding is trained on monolingual data for 5 iterations and used as an initialization for the RNN model.

2.5.4 AYLIEN_MULTILINGUAL **(Hokamp et al., 2019)**

The Aylien research team built a Multilingual NMT system which is trained on all WMT2019 language pairs in all directions, then fine-tuned for a small number of iterations on Gujarati-English data only, including some self-backtranslated data.

2.5.5 BAIDU **(Sun et al., 2019)**

Baidu systems are based on the Transformer architecture with several improvements. Data selection, back translation, data augmentation, knowledge distillation, domain adaptation, model ensemble and re-ranking are employed and proven effective in our experiments.

Team	Institution
AFRL	Air Force Research Laboratory (Gwinnup et al., 2019)
APERTIUM-FIN-ENG	Apertium (Pirinen, 2019)
APPRENTICE-C	Apprentice (Li and Specia, 2019)
AYLIEN_MULTILINGUAL	Aylien Ltd. (Hokamp et al., 2019)
BAIDU	Baidu (Sun et al., 2019)
BTRANS	(no associated paper)
BASELINE-RE-RERANK	(no associated paper)
CAIRE	(Liu et al., 2019)
CUNI	Charles University (Popel et al., 2019; Kocmi and Bojar, 2019) and (Kvapilíková et al., 2019)
DBMS-KU	Kumamoto University, Telkom University, Indonesian Institute of Sciences (Budiwati et al., 2019)
DFKI-NMT	DFKI (Zhang and van Genabith, 2019)
ETRANSLATION	eTranslation (Oravecz et al., 2019)
FACEBOOK FAIR	Facebook AI Research (Ng et al., 2019)
GTCOM	GTCOM (Bei et al., 2019)
HELSINKI NLP	University of Helsinki (Talman et al., 2019)
IIITH-MT	IIIT Hyderabad (Goyal and Sharma, 2019)
IITP	IIT Patna (Sen et al., 2019)
JHU	Johns Hopkins University (Marchisio et al., 2019)
JUMT	(no associated paper)
JU_SAARLAND	University of Saarland (Mondal et al., 2019)
KSAI	Kingsoft AI (Guo et al., 2019)
KYOTO UNIVERSITY	University of Kyoto (Cromieres and Kurohashi, 2019)
LINGUA CUSTODIA	Lingua Custodia (Burlot, 2019)
LIUM	LIUM (Bougares et al., 2019)
LMU-NMT	LMU Munich (Stojanovski and Fraser, 2019; Stojanovski et al., 2019)
MLLP-UPV	MLLP, Technical University of Valencia (Iranzo-Sánchez et al., 2019)
MS TRANSLATOR	Microsoft Translator (Junczys-Dowmunt, 2019)
MSRA	Microsoft Research Asia (Xia et al., 2019)
NIUTRANS	Northeastern University / NiuTrans Co., Ltd. (Li et al., 2019a)
NICT	National Institute of Information and Communications Technology (Dabre et al., 2019; Marie et al., 2019b)
NRC	National Research Council of Canada (Littell et al., 2019)
PARFDA	Boğaziçi University (Biçici, 2019)
PROMT-NMT	PROMT LLC (Molchanov, 2019)
RUG	University of Groningen (Toral et al., 2019)
RWTH AACHEN	RWTH Aachen (Rosendahl et al., 2019)
TALP_UPC_2019	TALP Research Center, Universitat Politècnica de Catalunya (Casas et al., 2019)
TARTUNLP-C	University of Tartu (Tättar et al., 2019)
TILDE-NC-NMT	Tilde (Pinnis et al., 2019)
UALACANT	Universitat d'Alacant (Sánchez-Cartagena et al., 2019)
UCAM	University of Cambridge (Stahlberg et al., 2019)
UDS-DFKI	Saarland University, DFKI (España-Bonet and Ruiter, 2019)
UEDIN	University of Edinburgh (Bawden et al., 2019a)
UMD	University of Maryland (Briakou and Carpuat, 2019)
USTC-MCC	(no associated paper)
USYD	University of Sydney (Ding and Tao, 2019)
XZL-NMT	(no associated paper)

Table 5: 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.

2.5.6 BTRANS

Unfortunately, no details are available for this system.

2.5.7 BASELINE-RE-RERANK (no associated paper)

BASELINE-RE-RERANK is a standard Transformer, with corpus filtering, pre-processing, post-processing, averaging and ensembling as well as n-best list reranking.

2.5.8 CAIRE (Liu et al., 2019)

CAIRE is a hybrid system that took part only in the unsupervised track. The system builds upon phrase-based MT and a pre-trained language model, combining word-level and subword-level NMT. A series of pre-processing and post-processing steps improves the performance, e.g. placeholders for numbers and dates, recasing and quotes normalization.

2.5.9 Charles University (CUNI) Systems

CUNI-T2T-TRANSFER (Kocmi and Bojar, 2019) are Transformer neural machine translation systems (as implemented in Tensor2tensor) for Kazakh↔English, Gujarati↔English. CUNI-T2T-TRANSFER focused on transfer learning from a high-resource language pair (Russian-English and Czech-English, respectively) followed by iterative back-translation.

CUNI-DOCTRANSFORMER-T2T2019 and CUNI-TRANSFORMER-T2T2019 (Popel et al., 2019) are trained in the T2T framework following the last year submission (Popel, 2018), but training on WMT19 document-level parallel and monoliongual data. During decoding, each document is split into overlapping multi-sentence segments, where only the "middle" sentences in each segment are used for the final translation. CUNI-TRANSFORMER-T2T2019 is the same system as CUNI-DOCTRANSFORMER-T2T2019, just applied on separate sentences during decoding.

CUNI-DOCTRANSFORMER-MARIAN (Popel et al., 2019) is a Transformer model as implemented in Marian and trained in a context-aware ("document-level") fashion. The training started with the same technique as the last year's submission but it was finetuned on document-level parallel and monolingual data by translating triples of adjacent sentences at once. If possible, only the middle sentence was considered for the final translation hypothesis, otherwise shorter context of two sentences or just a single sentence was used.

CUNI-TRANSFORMER-T2T2018 (Popel, 2018) is the exact same system as used last year.

CUNI-TRANSFORMER-MARIAN (Popel et al., 2019) is a "reimplementation" of the last year's system (Popel, 2018) in Marian (Junczys-Dowmunt et al., 2018).

CUNI-UNSUPERVISED-NER-POST (Kvapilíková et al., 2019) follows the strategy of Artetxe et al. (2018), creating a seed phrase-based system where the phrase table is initialized from cross-lingual embedding mappings trained on monolingual data, followed by a neural machine translation system trained on synthetic parallel corpus. The synthetic corpus is produced by the seed phrase-based MT system or by a such a model refined through iterative back-translation. CUNI-UNSUPERVISED-NER-POST further focuses on the handling of named entities, i.e. the part of vocabulary where the cross-lingual embedding mapping suffer most.

2.5.10 DBMS-KU (Budiwati et al., 2019)

The system DBMS-KU INTERPOLATION uses Linear Interpolation and Fillup Interpolation method with different language models, i.e., 3-gram and 5-gram. It combines a direct phrase table with pivot phrase table, pivoting through the Russian language.

2.5.11 DFKI-NMT (Zhang and van Genabith, 2019)

DFKI-NMT is a Transformer model trained using various techniques including data selection (using custom Transformer-based language models), back-translation and in-domain fine-tuning.

2.5.12 EN-DE-TASK

Unfortunately, no details are available for this system.

2.5.13 ETRANSLATION (Oravecz et al., 2019)

ETRANSLATION **En-De** ETRANSLATION's En-De system is an ensemble of 3 base Transformers and a Transformer-type language model, trained from all available parallel data (cleaned up and filtered with dual conditional cross-entropy filtering) and with additional back-translated data generated

from monolingual news. Each Transformer model is fine tuned on previous years' test sets.

ETRANSLATION **Fr-De** The Fr-De system is an ensemble of 2 big Transformers (with size 8192 FFN layers). Back-translation data was selected using topic modelling techniques to tune the model towards the domain defined in the task.

ETRANSLATION **En-Lt** The En-Lt system is an ensemble of 2 big Transformers (as for Fr-De) and a Transformer type language model. The training data contains the Rapid corpus and the news domain back-translated data sets 2 times oversampled.

ETRANSLATION **Ru-En** The Ru-En system is a single base Transformer trained only on true parallel data (including ParaCrawl but excluding the UN corpus) filtered in the same way as in the other submissions and fine tuned on previous test sets.

2.5.14 FACEBOOK FAIR (Ng et al., 2019)

Facebook FAIR system is a pure sentence level system, it is an ensemble of 3 Big Transformer models with FFN layers of size 8192. Trained on the mix of bitext and back-translated newscrawl data, oversampling was used to keep the effective ratio of bitext and back-translated data the same. Sampling from an ensemble of 3 models trained on bitext only was used to generate back-translations. The models were fine-tuned on in-domain data and a final noisy channel reranking was applied. All the training data (bitext and monolingual) was cleaned using langid filtering.

2.5.15 FRANK-S-MT

Unfortunately, no details are available for this system.

2.5.16 GTCOM (Bei et al., 2019)

GTCOM's systems (sysNameGTCOM-Primary) mainly focus on backtranslation, knowledge distillation and reranking to build a competitive model with transformer architecture. Also, the language model is applied to filter monolingual data, back-translated data and parallel data. The techniques for data filtering include filtering by rules, language models. Furthermore, they apply knowledge distillation techniques and right-to-left (R2L) reranking.

2.5.17 HELSINKI NLP (Talman et al., 2019)

HELSINKI NLP is a Transformer (Vaswani et al., 2017) style model implemented in OpenNMT-py using a variety of corpus filtering techniques, truecasing, BPE (Sennrich et al., 2016), back-translation, ensembling and fine-tuning for domain adaptation.

2.5.18 IIITH-MT (Goyal and Sharma, 2019)

IIITH-MT for Gujarati-English first experimented with attention-based LSTM encoder-decoder architecture, but later found the results to be more promising by using Transformer architecture. The paper documents that with Hindi-English as an assisting language pair in a joint training, the multilingual system obtains significant BLEU improvements for a low resource language pair like Gujarati-English.

2.5.19 IITP (Sen et al., 2019)

IITP-MT is a Transformer based NMT system trained using original parallel corpus and synthetic parallel corpus obtained through backtranslation of monolingual data. All the experiments are performed at subword-level using BPE with 10K merge operations.

2.5.20 JHU (Marchisio et al., 2019)

JHU's English-German system is an ensemble of 2 Transformer base models, improved by filtered backtranslation with restricted sampling (like Edunov+ 2018), filtered ParaCrawl and CommonCrawl (Junczys-Dowmunt, 2018a), continued training on newstest15-18 (like JHU's submission to WMT18, Koehn et al., 2018), reranking with R2L models (like Sennrich et al., 2017 or Junczys-Dowmunt, 2018b) and fixing quotation marks to match the German style (as many other teams did).

English-German was the same, with a 3 Transformer base ensemble, no fixed quotation marks, and reranking additionally included a language model (inspired by Junczys-Dowmunt, 2018a).

2.5.21 JUMT (no associated paper)

For the training purpose, the preprocessed Lithuanian-English sentence pairs were fed to Moses toolkit (Koehn et al., 2007). This created an SMT translation model with Lithuanian as the source language and English as the target language. After that, the Lithuanian side of a parallel corpus of 2,00,000 Lithuanian-English sentence pairs was re-translated into English with the

SMT model. These 2,00,000 machine translated English sentences and the respective 2,00,000 gold standard Lithuanian sentences (from the Lithuanian-English sentence pairs) were given as input to a word embedding based NMT model. This resulted in the hybrid model submitted for manual evaluation.

2.5.22 JU_SAARLAND (Mondal et al., 2019)

The systems JU_SAARLAND and JU_SAARLAND_CLEAN_NUM_135_BPE used additional backtranslated data and were trained using phrase-based and BPE-based attention models.

2.5.23 KSAI (Guo et al., 2019)

Kingsoft's submissions were based on various NMT architectures with Transformer as the baseline system. Several data filters and back-translation were used for data cleaning and data augmentation, respectively. Several advanced techniques were added to the baseline system such as Linear Combination and Layer Aggregation. Fine-tuning methods were applied to improve the in-domain translation quality. The final model was a system combination through multi-model ensembling and reranking, post-processed.

2.5.24 KYOTO UNIVERSITY (Cromieres and Kurohashi, 2019)

KYOTO UNIVERSITY used the now standard Transformer model (with 6 layers for each of encoder/decoder, hidden size of 1024, 16 attention heads, dropout of 0.3). Training data was carefully cleaned and the 2018 monolingual data was used through back-translation, as it turned out to be necessary for correctly translating recent news items. No ensemble translation was performed but a small BLEU improvement was obtained by taking a "majority vote" on the final translations for different checkpoints.

2.5.25 LINGUA CUSTODIA (Burlot, 2019)

The German-to-French system LINGUA-CUSTODIA-PRIMARY is an ensemble of eight Transformer *base* models, fine-tuned on monolingual news data back-translated with constrained decoding for specific terminology control.

2.5.26 LIUM (Bougares et al., 2019)

LIUM introduced two new translation directions involving two European languages: French and German. The training data was created by cross-matching the training data from previous WMT shared tasks. Development and test sets have been manually created from news articles Focusing on EU elections topics. LIUM participated in both directions for German-French language pairs. LIUM systems are based on the self-attentional Transformer networks using "small" and "big" architectures. We also used monolingual data selection and synthetic data through backtranslation.

2.5.27 LMU-NMT

LMU Munich provided two systems.

LMU-NMT (Stojanovski and Fraser, 2019) The LMU Munich system for En-De translation is based on a context-aware Transformer. We first train a baseline big Transformer on filtered ParaCrawl and an oversampled version of the remaining parallel data and then continue training with NewsCrawl backtranslations. We use the baseline to initialize the context-aware Transformer which uses fine-grained modeling of local and coarse-grained modeling of large context.

LMU-UNSUP (Stojanovski et al., 2019) The LMU Munich system for German-Czech translation is based on BWEs, cross-lingual LM, SMT and NMT, all trained in an unsupervised way. We train a cross-lingual Masked LM (Lample et al., 2019) and use it to initialize the NMT model. The NMT model is trained with denoising autoencoding and online backtranslation. We also include backtranslations from an unsupervised SMT. German data is compound-split and for NMT we further apply BPE splitting.

2.5.28 MLLP-UPV (Iranzo-Sánchez et al., 2019)

MLLP-UPV submitted systems for the German↔English and German↔French language pairs, participating in both directions of each pair. The systems are based on the Transformer architecture and make ample use of data filtering, synthetic data and domain adaptation through fine-tuning.

2.5.29 MS TRANSLATOR (Junczys-Dowmunt, 2019)

MS Translator systems (MICROSOFT-WMT19-SENT-DOC, MICROSOFT-WMT19-DOC-LEVEL and MICROSOFT-WMT19-SENT-LEVEL) explore the use of document-level context in large-scale

settings. We build 12-layer Transformer-Big systems: a) on the sentence-level, b) with large document-level context (training on full documents with up to 1024 subwords) and c) hybrid models via 2nd-pass decoding and ensembling. The models are trained on filtered parallel data, large amounts of back-translated documents and augmented fake and true parallel documents.

2.5.30 MSRA (Xia et al., 2019)

MSRA was submitted by Microsoft Research Asia. This system covers also the following sub-systems: MSRA.MADL, MSRA.MASS, MSRA.NAO and MSRA.SCA.

MSRA.MADL is based on Transformer (i.e., the standard transformer_big setting with 6 layers, embedding dimension 1024 and hidden state dimension 4096) and trained with multi-agent dual learning (Wang et al., 2019) scheme (briefly, MADL). The core idea of dual learning is to leverage the duality between the primal task (mapping from domain $\mathcal{X}$ to domain $\mathcal{Y}$) and dual task (mapping from domain $\mathcal{Y}$ to $\mathcal{X}$) to boost the performances of both tasks. MADL extends the dual learning framework by introducing multiple primal and dual models. It was integrated into the submitted system MSRA.MADL for German↔English and German↔French translations.

MSRA.SCA is a combination of Transformer network, back translation, knowledge distillation, soft contextual data augmentation (Zhu et al., 2019), and model ensembling. The Transformer big architecture is trained using soft contextual data augmentation to further enhance the performance. Following the above procedures, 5 different models are trained and ensembled for final submission.

MSRA.MASS is based on Transformer (i.e., the standard transformer_big setting with 6 layers, embedding dimension 1024 and hidden state dimension 4096) and pre-trained with MASS: masked sequence to sequence pre-training for language generation (Song et al., 2019). MASS leverages both monolingual and bilingual sentences for pre-training, where a segment of the source sentence is masked in the encoder side, and the decoder predicts this masked segment in the monolingual setting and predicts the whole target sentence in the bilingual setting. After pre-training, back-translation and ensemble/reranking are further leveraged to improve the accuracy of the system. MSRA.MASS handles Chinese→English and English↔Lithuanian translations in the submission

MSRA.NAO is a system whose architecture is obtained by neural architecture optimization (briefly, NAO; Luo et al., 2018). NAO leverages the power of a gradient-based method to conduct optimization and guide the creation of better neural architecture in a continuous and more compact space given the historically observed architectures and their performances. The search space includes self attention, convolutional networks, LSTMs, etc. It was applied in English↔Finnish translations in the submitted systems.

2.5.31 NIUTRANS providing the system NEU (Li et al., 2019a)

The NIUTRANS submissions are based on Deep-Transformer-DLCL and its variants, we used back-translation with beam search and sampling methods for data augmentation. Iterative ensemble knowledge distillation was employed to enhance single systems by various teachers. Ensembling and reranking facilitated further system combination.

2.5.32 NICT

NICT (Dabre et al., 2019) submitted supervised neural machine translation (NMT) systems developed for the news translation task for Kazakh↔English, Gujarati↔English, Chinese↔English, and English→Finnish translation directions.

NICT focused on leveraging multilingual transfer learning and back-translation for the extremely low-resource language pairs: Kazakh↔English and Gujarati↔English translation. For the Chinese↔English translation, back-translation, fine-tuning, and model ensembling were found to work the best. For English→Finnish, NICT submission from WMT18 remains a strong baseline despite the increase in parallel corpora for this year's task.

NICT (Marie et al., 2019b) submitted also an unsupervised neural machine translation system developed for the news translation task for German→Czech translation direction, focussing on language model pre-training, n-best list reranking, fine-tuning, and model ensembling technolo-

gies. The final primary submission to this task is the result of a simple combination of the unsupervised neural and statistical machine translation systems.

2.5.33 NRC (Littell et al., 2019)

The National Research Council Canada (NRC-CNRC) Kazakh-English news translation system is a multi-source, multi-encoder NMT system that takes Russian as the additional source. The constrained Kazakh-Russian parallel corpora is used to train NMT systems for "cross-translation" of resources between the languages, and the final Kazakh/Russian-to-English system is trained on a combination of genuine, back-translated, and cross-translated synthetic data. The submitted model is a partially trained single run system.

2.5.34 PARFDA (Biçici, 2019)

Biçici (2019) reports on the use of parfda system, Moses, KenLM, NPLM, and PRO, including the coverage of the test sets and the upper bounds on the translation results using the constrained resources.

2.5.35 PROMT-NMT (Molchanov, 2019)

This is an unconstrained, transformer-based single system, built using Marian and using BPE.

2.5.36 RUG

RUG_KKEN_MORFESSOR (Toral et al., 2019) uses (i) unsupervised morphological segmentation given the agglutinative nature of Kazakh, (ii) data from an additional language (Russian), given the scarcity of English–Kazakh data and (iii) synthetic data for the source language filtered using language-independent sentence similarity.

RUG_ENKK_BPE (Toral et al., 2019) uses data from an additional language (Russian), given the scarcity of English–Kazakh data and synthetic data (for both source and target languages) filtered using language-independent sentence similarity.

2.5.37 RWTH AACHEN (Rosendahl et al., 2019)

The systems by RWTH AACHEN are all based on Transformer architecture and aside from careful corpus filtering and fine tuning, they experiment with different types of subword units.

For English-German, no gains over the last year setup are observed. Small improvements are reached in Chinese-English. The highest gain of 11.1 BLEU is obtained for Kazakh-English, also thanks to transfer learning techniques.

2.5.38 TALP_UPC_2019_KKEN and TALP_UPC_2019_ENKK (Casas et al., 2019)

The TALP-UPC system was trained on a combination of the original Kazakh-English data (oversampled 3x) together with synthetic corpora obtained by translating with a BPE-based Moses the Russian side of the Kazakh-Russian data to English for the en-kk direction, and the Russian side of the English-Russian data to Kazakh for the kk-en direction. For the final systems, a custom model consisting in a self-attention Transformer decoder that learns joint source-target representations (with BPE tokenization) was used, implemented on the fairseq library.

2.5.39 TARTUNLP-C (Tättar et al., 2019)

TARTUNLP-C is a multilingual multi-domain neural machine translation, achieved by specifying the output language and domain via input word features (factors). The system was trained using all the parallel data for latin alphabet languages and used self-attention (Transformer) as the base architecture.

2.5.40 TILDE-NC-NMT and TILDE-NC-NMT (Pinnis et al., 2019)

Tilde developed both constrained and unconstrained NMT systems for English-Lithuanian and Lithuanian-English using the Marian toolkit. All systems feature ensembles of four to five transformer models that were trained using the quasi-hyperbolic Adam optimiser (Ma and Yarats, 2018). Data for the systems were prepared using TildeMT filtering (Pinnis, 2018) and preprocessing (Pinnis et al., 2018) methods. For unconstrained systems, data were additionally filtered using dual conditional cross-entropy filtering (Junczys-Dowmunt, 2018a). All systems were trained using iterative back-translation (Rikters, 2018) and feature synthetic data that allows training NMT systems to support handling of unknown phenomena (Pinnis et al., 2017). During translation, automatic named entity and non-translatable phrase post-editing were performed. For constrained systems, named entities and non-translatable phrase lists were extracted from the parallel training data. For unconstrained systems,

WikiData[8] was used to acquire bilingual lists of named entities.

2.5.41 Universitat d'Alacant

UALACANT-NMT **(Sánchez-Cartagena et al., 2019)** is an ensemble of two RNN and two transformer models. They were trained on a combination of genuine parallel data, synthetic data generated by means of pivot backtranslation (from the available English-Russian and Kazakh-Russian parallel data) and backtranslated monolingual data. The Kazakh text was morphologically segmented with Apertium.

UALACANT-NMT+RBMT **(Sánchez-Cartagena et al., 2019)** is an ensemble of two RNN and two Transformer models. They were trained on a combination of genuine parallel data, synthetic data generated by means of pivot backtranslation (from the available English-Russian and Kazakh-Russian parallel data) and backtranslated monolingual data. The Kazakh text was morphologically segmented with Apertium. The RNN models were multi-source models with two inputs: the original SL text and its translation with the Apertium RBMT English-Kazakh system.

2.5.42 UCAM (Stahlberg et al., 2019)

The Cambridge University Engineering Department's entry to the WMT19 evaluation campaign focuses on fine-tuning and language modelling. Fine-tuning on former WMT test sets is regularized with elastic weight consolidation (Kirkpatrick et al., 2017). Language models are used on both the sentence-level and the document-level, with a modified Transformer architecture for document-level language modelling. An SMT system is integrated via a minimum Bayes-risk formulation (Stahlberg et al., 2017).

2.5.43 UDS-DFKI (España-Bonet and Ruiter, 2019)

The UdS-DFKI English→German system uses a standard Transformer architecture where data is enriched with coreference information gathered at document level. The training is still done at the sentence level.

The English↔Gujarati systems are phrase-based SMT systems enriched with parallel sentences extracted from comparable corpora with a self-supervised NMT system. In this case, also back-translations are used.

2.5.44 UEDIN (Bawden et al., 2019a)

The UEDIN systems are supervised NMT systems based on the transformer architecture and trained using Marian (Junczys-Dowmunt et al., 2018). For English↔Gujarati, synthetic parallel data from two sources, backtranslation and pivoting through Hindi, is produced using unsupervised and semi-supervised NMT models, pre-trained using a cross-lingual language objective (Lample and Conneau, 2019) For German→English, the impact of vast amounts of back-translated training data on translation quality is studied, and some additional insights are gained over (Edunov et al., 2018). Towards the end of training, for German→English and Chinese↔English, the mini-batch size was increased up to fifty-fold by delaying gradient updates (Bogoychev et al., 2018) as an alternative to learning rate cooldown (Smith, 2018). For Chinese↔English, a comparison of different segmentation strategies showed that character-based decoding was superior to the translation of subwords when translating into Chinese. Pre-processing strategies were also investigated for English→Czech, showing that pre-processing can be simplified without loss to MT quality.

UEDIN's main findings on the Chinese↔English translation task are that character-level model on the Chinese side can be used when translating into Chinese to improve the BLEU score. The same does not hold when translating from Chinese.

2.5.45 UMD (Briakou and Carpuat, 2019)

UMD NMT models are Sequence-2-Sequence attentional with Long-Short Term Memory units; words are segmented using BPEs jointly learned on the concatenation of Turkish and Kazakh data. The submitted model is an ensemble obtained by averaging the output distributions of 4 models trained on Kazakh, Turkish and back-translated data using different random seeds.

2.5.46 UNSUPERVISED-6929 and UNSUPERVISED-6935

Unfortunately, no details are available for these systems.

[8] www.wikidata.org

2.5.47 USTC-MCC (no associated paper)

USTC-MCC is a Transformer model implemented in Fairseq-py. Tokenization and BPE were used and the training data were augmented with back-translation.

2.5.48 USYD (Ding and Tao, 2019)

The University of Sydney's system is based on the self attentional Transformer networks, into which they integrated the most recent effective strategies from academic research (e.g., BPE, back translation, multi-features data selection, data augmentation, greedy model ensemble, reranking, ConMBR system combination, and post-processing). Furthermore, they proposed a novel augmentation method Cycle Translation and a data mixture strategy Big/Small parallel construction to entirely exploit the synthetic corpus.

2.5.49 XZL-NMT (no associated paper)

XZL-NMT is an ensembled Transformer model as implemented in Marian, using Moses tokenizer and subword units.

2.6 Submission Summary

An overview of techniques used in the submitted systems was obtained in a poll. The full details are available on-line.[9] Including manually entered data rows, we had more than 60 responses, some of which describe more MT systems at once.

Overall, most of the submitted systems were standard bilingual MT systems, optimized to translate one language pair, even in the case when data from other languages are used to support this pair. Truly multilingual systems were TARTUNLP-C covering 7 of the tested language pairs, DBMS-KU INTERPOLATION (bidirectional Kazakh-English) and AYLIEN_MT_MULTILINGUAL which was unfortunately tested only on the very low-resource Gujarati-English and not all the language pairs it covers. In the highly competitive task of news translation, these systems ended up on lower ranks, so aiming at multi-linguality seems rather as a distraction, except for supporting low-resource languages.

As already in the previous year, the Transformer architecture (Vaswani et al., 2017) dominates with more than 80% of submissions[10] reporting to include it. Some diversity is seen at least in the actual implementation of the model, with Marian (Junczys-Dowmunt et al., 2018) being by far the most popular (more than 30%), followed by fairseq (18%), OpenNMT-py (16%) and Tensor2tensor and Sockeye (14% each). Phrase-based MT (primarily Moses, Koehn et al., 2007) is still often in use, with 15–25% submissions using it in some way.

Feature	#	[%]
Dropout	42	69
Back-translation	39	64
Ensembling	37	61
Careful corpus filtering	35	57
Tied source and target word embeddings	24	39
Fine-tuning for domain adaptation	22	36
Back-translation more than once	20	33
Averaging	17	28
Oversampling	14	23
Extra languages used (e.g. some form of pivoting or multi-lingual training)	12	20
Pre-trained model parts (e.g. word embeddings)	10	16
Total	61	100

Table 6: Model and training features frequently reported for submitted systems.

Subword processing is very frequent: BPE (Sennrich et al., 2016) taking the lead (two thirds) and SentencePiece (Kudo and Richardson, 2018) following (a quarter of submissions). More than 90% of submissions use tokenization (Moses tokenizer being used in 40% of cases) before subword splitting while more language-specific tools such as morphological segmenters are rare. Unicode characters were used only exceptionally (4 mentions) and with rather experimental systems, except for UEDIN, see Section 2.5.44.

More than 40% of submissions used language identification to clean the provided training data. Truecasing or recasing was also quite popular.

Common NMT model and training features are listed in Table 6, documenting that back-translation, ensembling and corpus filtering are a must.

3 Human Evaluation

A human evaluation campaign is run each year to assess translation quality and to determine the final ranking of systems taking part in the competition.

[9] https://tinyurl.com/wmt19-systems-descr-summary

[10] The percentages are indicative only. They are based on the total number of responses in the poll, with only an inexact correspondence to the number of evaluated primary submissions.

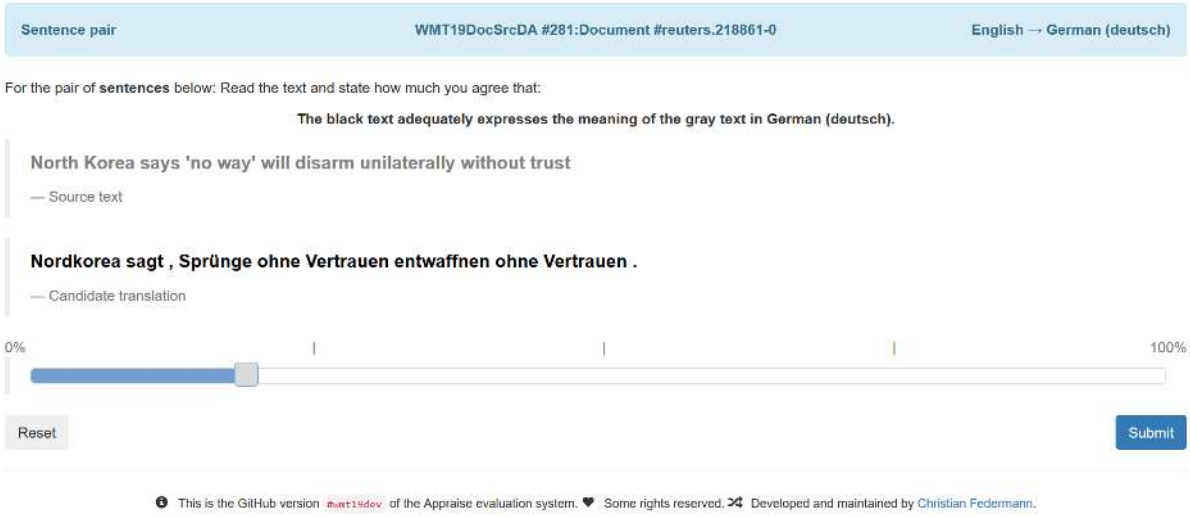

Figure 3: Screen shot of segment-rating portion of document-level direct assessment in the Appraise interface for an example English to German assessment from the human evaluation campaign. The annotator is presented with the machine translation output segment randomly selected from competing systems (anonymized) and is asked to rate the translation on a sliding scale.

This section describes how preparation of evaluation data, collection of human assessments, and computation of the official results of the shared task was carried out this year.

3.1 Direct Assessment

Work on evaluation over the past few years has provided fresh insight into ways to collect *direct assessments* (DA) of machine translation quality (Graham et al., 2013, 2014, 2016), and three years ago the evaluation campaign included parallel assessment of a subset of News task language pairs evaluated with *relative ranking* (RR) and DA. DA has some clear advantages over RR, namely the evaluation of absolute translation quality and the ability to carry out evaluations through quality controlled crowd-sourcing. As established in 2016 (Bojar et al., 2016), DA results (via crowd-sourcing) and RR results (produced by researchers) correlate strongly, with Pearson correlation ranging from 0.920 to 0.997 across several source languages into English and at 0.975 for English-to-Russian (the only pair evaluated out-of-English). Since 2017, we have thus employed DA for evaluation of systems taking part in the news task and do so again this year.

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. No sentence or document length restriction is applied during manual evaluation.

3.2 Styles of Direct Assessment Tested in WMT19

In previous year's evaluation translated segments for all language pairs were evaluated independent of the wider document context. However, since recent MT evaluations address the question of comparison of system and human performance, evaluation within document context has become more relevant (Läubli et al., 2018; Toral et al., 2018). Therefore, for a selection of language pairs, human evaluation was carried out within the document context. We denote the two options "+DC" (with document context) and "−DC" (without document context) in the following.

Additionally in past years, test data included text that was created in the opposite direction to testing, in order to achieve a larger test set with limited resources. Inclusion of test data has been shown to introduce inaccuracies in evaluations particularly in terms of BLEU scores however (Graham et al., 2019b) and for this reason, this year we only test systems on data that was originally written in the source language.

In previous years we have employed only monolingual human evaluation (denoted "M" in the following for official results. Last year we trialled source-based evaluation for English to Czech translation, i.e. a bilingual configuration ("B") in which the human assessor is shown the source input and system output only (with no reference

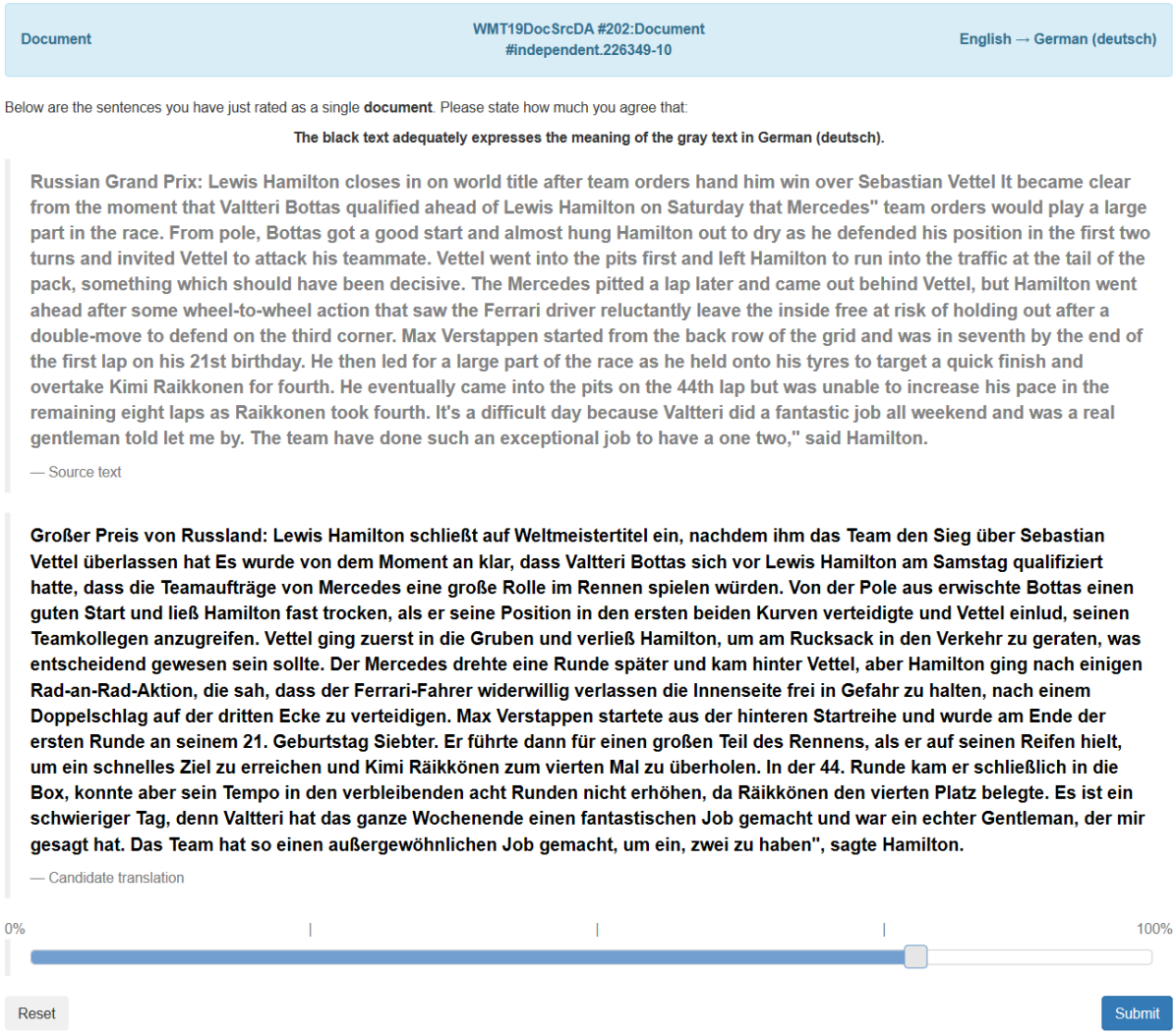

Figure 4: Screen shot of document-rating portion of document-level direct assessment in the Appraise interface for an example English to German assessment from the human evaluation campaign. The annotator is presented with the machine translation output document randomly selected from competing systems (anonymized) and is asked to rate the translation on a sliding scale.

translation shown). This approach has the advantage of freeing up the human-generated reference translation so that it can be included in the evaluation as another system and provide an estimate of human performance. Since we would like to restrict human assessors to only evaluate translation *into* their native language, we restricted 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, native English speakers with high fluency in non-English languages. A summary of the human evaluation configurations run this year in the news task is provided in Table 7, where configurations that correspond to official results are highlighted in bold.

The style of official evaluation used in the past recent years of WMT corresponds to M SR−DC (Segment Rating without Document Context) i.e. evaluating individual segments against the reference translation and independently of each other.

For language pairs for which our original style SR−DC evaluation was run this year, the SR−DC configuration was kept as the source of the official results with additional configurations provided for the purpose of comparison. For the remaining language pairs, official results are based on the SR+DC evaluation, i.e. the assessment of individual segments which are nevertheless provided in their natural order as they appear in the document. Fully document-level evaluation (DR+DC)

	Doc Rating + Doc Context (DR+DC)	Seg Rating + Doc Context (SR+DC)	Seg Rating − Doc Context (SR−DC)
de-cs			**M**
de-fr			**M**
fr-de			**M**
de-en	M	**M**	
en-cs	B	**B**	B
en-de	B	**B**	
en-fi	B	**B**	
en-gu	B	**B**	
en-kk	B	**B**	
en-lt	B	**B**	
en-ru	B	**B**	
en-zh	B	**B**	
fi-en			**M**
gu-en			**M**
kk-en			**M**
lt-en			**M**
ru-en			**M**
zh-en	M	M	**M**

Table 7: Summary of human evaluation configurations; M denotes reference-based/monolingual human evaluation in which the machine translation output was compared to human-generated reference; B denotes bilingual/source-based evaluation where the human annotators evaluated MT output by reading the source language input only (no reference translation present); configurations comprising official results highlighted in bold.

as trialled this 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. (2019b).

In order to maximize the number of human annotations collected while minimizing the amount of reading required by a given human assessor, we combined two evaluation configurations, Document Rating + Document Context (DR+DC) and Segment Rating + Document Context (SR+DC), shown in Table 7 and ran them as a single task. In this configuration, human annotators were shown each segment of a given document (produced by a single MT system) in original sequential order and the human assessor rated each segment in turn. Figure 3 shows a screenshot of this part of the annotation process. This was followed by a screen where the human assessor rated the entire document as a whole comprising the most recently rated segments. Figure 4 shows this later part of the same evaluation set-up. Subsequently when sufficient data is collected, SR+DC results are arrived at by combining ratings attributed to segments, while DR+DC results are a combination of document ratings.

For some language pairs the standard configuration from past years in which segments are evaluated in isolation from the wider document context, which we call Segment Rating − Document Context (SR−DC) and a screenshot of this configuration is shown in Figure 5.

As in previous years, the standard SR−DC annotation is organized into "HITs" (following the Mechanical Turk's term "human intelligence task"), each containing 100 such screens and requiring about half an hour to finish. For the additional configuration that included both DR+DC and SR+DC, HITs were simply made up of a random sample of machine translated *documents* as opposed to segments.

3.3 Evaluation Campaign Overview

In terms of the News translation task manual evaluation, a total of 263 individual researcher accounts were involved, and 766 turker accounts.[11] Researchers in the manual evaluation contributed judgments of 242,424 translations, while 487,674 translation assessment scores were submitted in total by the crowd, of which 224,046 were provided by workers who passed quality control.

Under ordinary circumstances, each assessed translation would correspond to a single individual scored segment. However, since distinct systems can produce the same output for a particular input sentence, in previous years we were often able to take advantage of this and use a single assessment for multiple systems. For example, last year we combined human assessment of identical translations produced by multiple systems and were able to get up to 17% saving in terms of evaluation resources. However, since our evaluation now includes document context, deduplication of system outputs was not possible for most of the configurations run this year.

3.4 Data Collection

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. Annotations are collected in an evaluation campaign that enlists the help of participants in the shared task. Each team is asked to contribute 8 hours annotation time, which we estimated at 16 100-translation HITs per primary system submitted. We continue to use the open-source Appraise[12] (Federmann, 2012) tool and Turkle2 for

[11] Numbers do not include the 1,005 workers on Mechanical Turk who did not pass quality control.

[12] https://github.com/cfedermann/Appraise

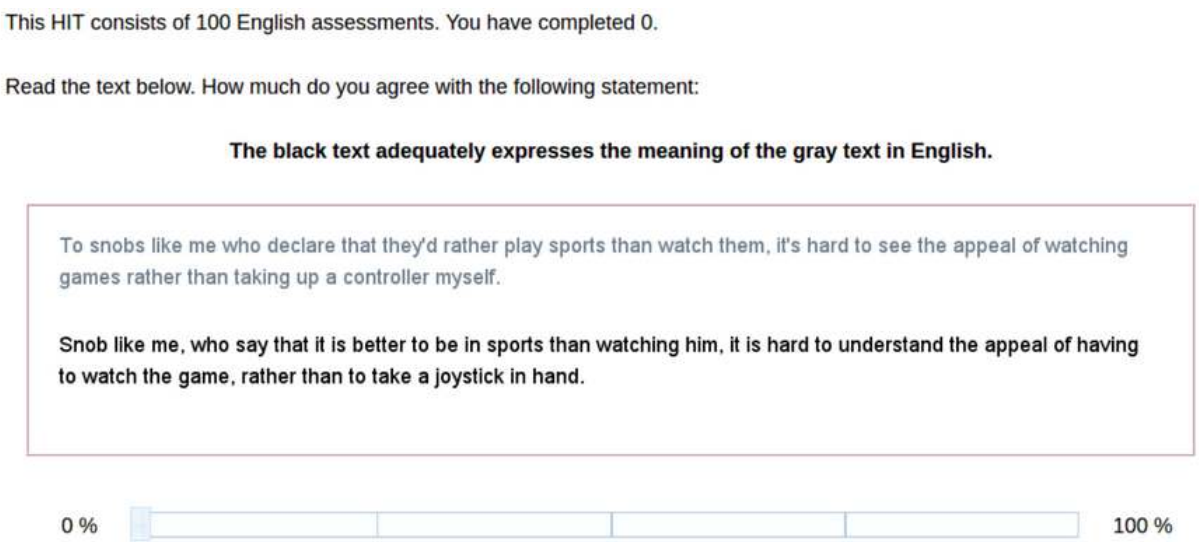

Figure 5: Screen shot of Direct Assessment as carried out by workers for the standard Segment Rating − Document Context (SR−DC) Human Evaluation Configuration.

our data collection, in addition to Amazon Mechanical Turk.[13] Table 8 shows total numbers of human assessments collected in WMT19 contributing to final scores for systems.[14]

The effort that goes into the manual evaluation campaign each year is impressive, and we are grateful to all participating individuals and teams. We believe that human annotation provides the best decision basis for evaluation of machine translation output and it is great to see continued contributions on this large scale.

3.5 Crowd Quality Control

In order to trial document-level evaluation, in addition to our standard segment-level human evaluation, we ran two additional evaluations combined into a single HIT structure. Firstly, we collected segment ratings with document context (SR+DC) and secondly document ratings with document context (DR+DC). We refer to our original segment-level evaluation where assessors are shown segments in isolation from the wider document context as segment rating − document context (SR−DC). We describe all three methods of ranking systems in detail below.

3.5.1 Standard DA HIT Structure (SR−DC)

In the standard DA HIT structure (Segment Rating − Document Context), three kinds of quality control translation pairs are employed as described in Table 9: we repeat pairs (expecting a similar judgment), damage MT outputs (expecting significantly worse scores) and use references instead of MT outputs (expecting high scores).

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 averaged out. Furthermore apart from quality con-

[13] https://www.mturk.com

[14] Number of systems for WMT19 includes ten "human" systems comprising human-generated reference translations used to provide human performance estimates.

Language Pair	Systems	Comps	Comps/Sys	Assessments	Assess/Sys
Chinese→English	15	–	–	20,199	1,346.6
German→English	17	–	–	39,556	2,326.8
Finnish→English	12	–	–	23,301	1,941.8
Gujarati→English	11	–	–	17,147	1,558.8
Kazakh→English	11	–	–	18,339	1,667.2
Lithuanian→English	11	–	–	18,807	1,709.7
Russian→English	14	–	–	27,836	1,988.3
English→Chinese	13	–	–	28,801	2,215.5
English→Czech	12	–	–	29,207	2,433.9
English→German	23	–	–	49,535	2,153.7
English→Finnish	13	–	–	22,310	1,716.2
English→Gujarati	12	–	–	11,223	935.2
English→Kazakh	13	–	–	15,039	1,156.8
English→Lithuanian	13	–	–	14,069	1,082.2
English→Russian	13	–	–	24,441	1,880.1
German→Czech	11	–	–	16,900	1,536.4
German→French	11	–	–	6,700	609.1
French→German	10	–	–	4,000	400.0
Total Appraise	112	–	–	194,625	1,737.7
Total MTurk	76	–	–	144,986	1,907.7
Total Turkle	47	–	–	47,799	1,017.0
Total WMT19	243	–	–	387,410	1,594.3
WMT18	150	–	–	302,489	2,016.6
WMT17	153	–	–	307,707	2,011.2
WMT16	138	569,287	4,125.2	284,644	2,062.6
WMT15	131	542,732	4,143.0	271,366	2,071.5
WMT14	110	328,830	2,989.3	164,415	1,494.7
WMT13	148	942,840	6,370.5	471,420	3,185.3
WMT12	103	101,969	999.6	50,985	495.0
WMT11	133	63,045	474.0	31,522	237.0

Table 8: Amount of data collected in the WMT19 manual evaluation campaign (after removal of quality control items). The final eight rows report summary information from previous years of the workshop.

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.

trol items, HITs are constructed using translations sampled from the entire set of outputs for a given language pair.

3.5.2 Document-Level DA HIT Structure (SR+DC and DR+DC)

As mentioned previously, collection of segment-level ratings with document context (Segment Rating + Document Context) and document ratings with document context (Document Rating + Document Context) assessments were combined into a single evaluation set-up to save annotator time. This involved constructing HITs so that each sentence belonging to a given document (produced by a single MT system) were displayed to and rated by the human annotator before he/she was shown the same entire document again and asked to rate it.

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

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

2. Documents are then sampled at random (without replacement) and assigned to the current HIT until the current HIT comprises

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

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 Section 3.5.3 for details of bad reference generation;

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

3.5.3 Construction of Bad References

In all set-ups employed in the evaluation campaign, and 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 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	⌊ N/4 ⌋

3.6 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 filter crowd-sourced human assessors by how consistently they rate translations of known distinct quality using the bad reference pairs described previously. Quality filtering via bad reference pairs is especially important for the crowd-sourced portion of the manual evaluation. Due to the anonymous nature of crowd-sourcing, when collecting assessments of translations, it is likely to encounter workers who attempt to game the service, as well as submission of inconsistent evaluations and even robotic ones. We therefore employ DA's quality control mechanism to filter out low quality data, facilitated by the use of DA's analogue rating scale.[16]

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 who met our filtering requirement by showing a signif-

[16] As stated previously, this year we removed the requirement for volunteer researchers to annotate quality control items and this also removes the possibility to report agreement statistics for this group.

icantly 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.

Numbers in Table 10 of workers passing quality control criteria (A) varies across language pairs and this is in-line with passed DA evaluations. Language pairs were run in the following order on Mechanical Turk: fi-en, gu-en, kk-en, lt-en ru-en, zh-en, de-en. We observe that the amount of low quality data we received (with one exception at the beginning) steadily decreases as data collection proceeded from (100−31=) 69% low quality data for fi-en to (100−71=) 29% for de-en, the last language pair to be evaluated. This is likely due to the active rejection of low quality HITs and word spreading among unreliable workers to avoid our HITs. The assessors were least reliable for gu-en, with only 60 out of 301 workers passing the quality control. We removed the data from the non-reliable workers in all language pairs.

In terms of numbers of workers who passed quality control who also showed no significant difference in exact repeats of the same translation, the two document-level runs, zh-en and de-en, showed lower reliability than the original DA standard sentence-level set-up. Overall the reliability is still relatively high however with the lowest language pair being de-en still reaching 88% of workers showing no significant difference in scores for repeat assessment of the same translation. In sum, we confirmed this year again that the check on bad references is sufficient and not many more workers would be ruled out if we also demanded similar judgements for repeated inputs.

3.7 Producing the Human Ranking

The data belong to each individual human evaluation run were compiled individually to produce either one of our official system rankings or a ranking that we would like to compare with official rankings.

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. For rankings arrived at via segment ratings (SR−DC as well as 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 11). For rankings arrived at via document ratings (DR+DC), average standardized scores for individual documents belonging to a given system were then computed, before the final overall DA score for a given system was computed as the average of its document scores (Ave z in Table 11). Results are also reported for average scores for systems, computed in the same way but without any score standardization applied (Ave % in Table 11).

Tables 11, Tables 12 and 13 include the official results of the news task and Tables 14 and 15 include results for alternate human evaluation configurations.[17] Human performance estimates arrived at by evaluation of human-produced reference translations are denoted by "HUMAN" in all tables. Clusters are identified by grouping systems together according to which systems significantly outperform all others in lower ranking clusters, according to Wilcoxon rank-sum test. Appendix A shows the underlying head-to-head significance test official results for all pairs of systems.

3.8 Human Parity

In terms of human parity, as pointed out by Graham et al. (2019b), fully document-level evaluations incur the problem of low statistical power due to the reduced sample size of documents. The many ties in our DR+DC evaluation results cannot be used to draw conclusions of human parity with MT therefore. In addition, as highlighted by Toral et al. (2018), Läubli et al. (2018) and also us Bojar et al. (2018), a tie of human and machine in an evaluation of isolated segments cannot be used to draw conclusions of human parity. Given a wider context, human evaluators may draw different conclusions.[18]

Our SR+DC human evaluation configuration is an attempt to draw the right balance between making it possible to assess a sufficient sample size of translations but importantly keeping the docu-

[17] See Table 7 for human evaluation configuration details of each language pair

[18] The only setting where segment-level evaluation could serve in human-parity considerations would be when both humans and machines were translating isolated segments but this setting is not very interesting from the practical point of view.

Order		All	(A) Sig. Diff. Bad Ref.	(B) (A) & No Sig. Diff. Exact Rep.
1	Finnish→English	443	137 (31%)	135 (99%)
2	Gujarati→English	301	60 (20%)	59 (98%)
3	Kazakh→English	217	73 (34%)	70 (96%)
4	Lithuanian→English	233	90 (39%)	85 (94%)
5	Russian→English	321	137 (43%)	136 (99%)
6	Chinese→English	440	208 (47%)	186 (89%)
7	German→English	380	268 (71%)	236 (88%)
	Total	**1,706**	**766 (45%)**	**711 (93%)**

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; (B) those of (A) whose scores also showed no significant difference for exact repeats of the same translation. The language pairs were submitted for evaluation one after another in the reported order.

ment context available to human assessors, a configuration highlighted as suitable for human-parity investigations by Graham et al. (2019b) and already employed by Toral et al. (2018) (although our overall evaluation differs in other respects). According to the power analysis provided in Graham et al. (2019b), the sample size of translations evaluated in the set-up is large enough to safely conclude statistical ties between pairs of systems in our SR+DC configurations. In addition our evaluation meets all requirements included on the MT evaluation checklist of Graham et al. (2019b).

The results that can be relied upon for drawing conclusions of human parity therefore include the following from our SR+DC configurations:

✓ German to English: many systems are tied with human performance;

× English to Chinese: all systems are outperformed by the human translator;

× English to Czech: all systems are outperformed by the human translator;

× English to Finnish: all systems are outperformed by the human translator;

✓ English to German: Facebook-FAIR achieves super-human translation performance; several systems are tied with human performance;

× English to Gujarati: all systems are outperformed by the human translator;

× English to Kazakh: all systems are outperformed by the human translator;

× English to Lithuanian: all systems are outperformed by the human translator;

✓ English to Russian: Facebook-FAIR is tied with human performance.

Even with all our precautions, the indications of human parity should not be overvalued. For instance, the super-human performance observed with Facebook-FAIR on English to German is based on standardized scores (Ave z.). Without the standardization (Ave.), Facebook-FAIR is on par with the reference and two systems by Microsoft score higher. The same mismatch of Ave. and Ave. z happens for English-Czech within the second performance cluster and also a couple of times in German-English and other language pairs. This has happened in the past already but the English-German case seems to be the first one where the Wilcoxon test claims a significant difference.

3.9 Comparing the Different English-Czech Results

Table 16 reproduces English-to-Czech official SR+DC scores and the full-document DR+DC, to compare them with two additional runs of the bilingual SR−DC style, i.e. the exact same context-less setting used in source-based evaluation of en2cs in WMT18 where the quality of the reference has been significantly surpassed.

The results "SR−DC WMT" are based on 6,225 judgements (518 per system) collected by the same set of annotators as the official SR+DC scores and the "SR−DC Microsoft" are based on 21,918 judgements (1,826 per system) sponsored and carried out by Microsoft.

English→German

Ave.	Ave. z	System
90.3	0.347	Facebook-FAIR
93.0	0.311	Microsoft-WMT19-sent-doc
92.6	0.296	Microsoft-WMT19-doc-level
90.3	0.240	HUMAN
87.6	0.214	MSRA-MADL
88.7	0.213	UCAM
89.6	0.208	NEU
87.5	0.189	MLLP-UPV
87.5	0.130	eTranslation
86.8	0.119	dfki-nmt
84.2	0.094	online-B
86.6	0.094	Microsoft-WMT19-sent-level
87.3	0.081	JHU
84.4	0.077	Helsinki-NLP
84.2	0.038	online-Y
83.7	0.010	lmu-ctx-tf-single
84.1	0.001	PROMT-NMT
82.8	−0.072	online-A
82.7	−0.119	online-G
80.3	−0.129	UdS-DFKI
82.4	−0.132	TartuNLP-c
76.3	−0.400	online-X
43.3	−1.769	en-de-task

Gujarati→English

Ave.	Ave. z	System
64.8	0.210	NEU
61.7	0.126	UEDIN
59.4	0.100	GTCOM-Primary
60.8	0.090	CUNI-T2T-transfer
59.4	0.066	aylien-mt-multilingual
59.3	0.044	NICT
51.3	−0.189	online-G
50.9	−0.192	IITP-MT
48.0	−0.277	UdS-DFKI
47.4	−0.296	IIITH-MT
41.1	−0.598	Ju-Saarland

English→Gujarati

Ave.	Ave. z	System
73.1	0.701	HUMAN
72.2	0.663	online-B
66.8	0.597	GTCOM-Primary
60.2	0.318	MSRA
58.3	0.305	UEDIN
55.9	0.254	CUNI-T2T-transfer
52.7	−0.079	Ju-Saarland-clean-num-135-bpe
35.2	−0.458	IITP-MT
38.8	−0.465	NICT
39.1	−0.490	online-G
33.1	−0.502	online-X
33.2	−0.718	UdS-DFKI

Kazakh→English

Ave.	Ave. z	System
72.2	0.270	online-B
70.1	0.218	NEU
69.7	0.189	rug-morfessor
68.1	0.133	online-G
67.1	0.113	talp-upc-2019
67.0	0.092	NRC-CNRC
65.8	0.066	Frank-s-MT
65.6	0.064	NICT
64.5	0.003	CUNI-T2T-transfer
48.9	−0.477	UMD
32.1	−1.058	DBMS-KU

Lithuanian→English

Ave.	Ave. z	System
77.4	0.234	GTCOM-Primary
77.5	0.216	tilde-nc-nmt
77.0	0.213	NEU
76.4	0.206	MSRA-MASS
76.4	0.202	tilde-c-nmt
73.8	0.107	online-B
69.4	−0.056	online-A
69.2	−0.059	TartuNLP-c
62.8	−0.284	online-G
62.4	−0.337	JUMT
59.1	−0.396	online-X

German→English

Ave.	Ave. z	System
81.6	0.146	Facebook-FAIR
81.5	0.136	RWTH-Aachen
79.0	0.136	MSRA-MADL
79.9	0.121	online-B
79.0	0.086	JHU
80.1	0.067	MLLP-UPV
79.0	0.066	dfki-nmt
78.0	0.066	UCAM
76.6	0.050	online-A
78.4	0.039	NEU
79.0	0.027	HUMAN
77.4	0.011	uedin
77.9	0.009	online-Y
74.8	0.006	TartuNLP-c
72.9	−0.051	online-G
71.8	−0.128	PROMT-NMT
69.7	−0.192	online-X

English→Czech

Ave.	Ave. z	System
91.2	0.642	HUMAN
86.0	0.402	CUNI-DocTransformer-T2T
86.9	0.401	CUNI-Transformer-T2T-2018
85.4	0.388	CUNI-Transformer-T2T-2019
81.3	0.223	CUNI-DocTransformer-Marian
80.5	0.206	uedin
70.8	−0.156	online-Y
71.4	−0.195	TartuNLP-c
67.8	−0.300	online-G
68.0	−0.336	online-B
60.9	−0.594	online-A
59.3	−0.651	online-X

Finnish→English

Ave.	Ave. z	System
78.2	0.285	MSRA-NAO
77.8	0.265	online-Y
77.6	0.261	GTCOM-Primary
76.4	0.245	USYD
72.5	0.107	online-B
73.3	0.105	Helsinki-NLP
69.2	0.012	online-A
68.4	−0.044	online-G
68.0	−0.053	TartuNLP-c
67.3	−0.071	online-X
61.9	−0.209	parfda
53.3	−0.516	apertium-uc

English→Finnish

Ave.	Ave. z	System
94.8	1.007	HUMAN
82.6	0.586	GTCOM-Primary
80.2	0.570	MSRA-NAO
70.9	0.275	online-Y
65.8	0.199	NICT
65.7	0.09	Helsinki-NLP
63.1	0.072	online-G
63.0	0.037	online-B
54.5	−0.125	TartuNLP-c
48.3	−0.384	online-A
47.1	−0.398	online-X
47.9	−0.522	Helsinki-NLP-rule-based
16.9	−1.260	apertium-uc

English→Kazakh

Ave.	Ave. z	System
81.5	0.746	HUMAN
67.6	0.262	UAlacant-NMT
63.8	0.243	online-B
63.8	0.222	UAlacant-NMT-RBMT
63.3	0.126	NEU
63.3	0.108	MSRA
60.4	0.097	CUNI-T2T-transfer
61.7	0.078	online-G
55.2	−0.049	rug-bpe
49.0	−0.328	talp-upc-2019
41.4	−0.493	NICT
11.6	−1.395	DBMS-KU

English→Lithuanian

Ave.	Ave. z	System
90.5	1.017	HUMAN
72.8	0.388	tilde-nc-nmt
69.1	0.387	MSRA-MASS-uc
68.0	0.262	tilde-c-nmt
68.2	0.259	MSRA-MASS-c
67.7	0.155	GTCOM-Primary
62.7	0.036	eTranslation
59.6	−0.054	NEU
57.4	−0.061	online-B
47.8	−0.383	TartuNLP-c
38.4	−0.620	online-A
39.2	−0.666	online-X
32.6	−0.805	online-G

English→Russian

Ave.	Ave. z	System
89.5	0.536	HUMAN
88.5	0.506	Facebook-FAIR
83.6	0.332	USTC-MCC
82.0	0.279	online-G
80.4	0.269	online-B
79.0	0.223	NEU
80.2	0.219	PROMT-NMT
78.5	0.156	online-Y
71.7	−0.188	rerank-er
67.9	−0.268	online-A
68.8	−0.310	TartuNLP-u
62.1	−0.363	online-X
35.7	−1.270	NICT

English→Chinese

Ave.	Ave. z	System
82.5	0.368	HUMAN
83.0	0.306	KSAI
83.3	0.280	Baidu
80.5	0.209	NEU
80.3	0.052	online-A
79.9	0.042	xzl-nmt
79.0	0.017	UEDIN
77.8	0.009	BTRANS
76.9	0.000	NICT
74.6	−0.125	online-B
75.6	−0.218	online-Y
72.6	−0.262	online-G
69.5	−0.553	online-X

Russian→English

Ave.	Ave. z	System
81.4	0.156	Facebook-FAIR
80.7	0.134	online-G
80.4	0.122	eTranslation
80.1	0.121	online-B
81.4	0.115	NEU
80.4	0.102	MSRA-SCA
79.8	0.084	rerank-re
79.2	0.076	online-Y
79.0	0.029	online-A
76.8	0.012	afrl-syscomb19
76.8	−0.039	afrl-ewc
76.2	−0.040	TartuNLP-u
74.5	−0.097	online-X
69.3	−0.303	NICT

Chinese→English

Ave.	Ave. z	System
83.6	0.295	Baidu
82.7	0.266	KSAI
81.7	0.203	MSRA-MASS
81.5	0.195	MSRA-MASS
81.5	0.193	NEU
80.6	0.186	BTRANS
80.7	0.161	online-B
79.2	0.103	BTRANS-ensemble
77.9	0.054	UEDIN
78.0	0.049	online-Y
77.4	0.001	NICT
75.3	−0.065	online-A
72.4	−0.202	online-G
66.9	−0.483	online-X
56.4	−0.957	Apprentice-c

Table 11: Official results of WMT19 News Translation Task. 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→Czech

Ave.	Ave. z	System
63.9	0.426	online-Y
62.7	0.386	online-B
61.4	0.367	NICT
59.8	0.319	online-G
55.7	0.179	NEU-KingSoft
54.4	0.134	online-A
47.8	−0.099	lmu-unsup-nmt
46.6	−0.165	CUNI-Unsupervised-NER-post
41.7	−0.328	Unsupervised-6929
39.1	−0.405	Unsupervised-6935
28.4	−0.807	CAiRE

Table 12: Official results of WMT19 German to Czech Unsupervised News Translation Task. 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 (in particular the use of parallel training data).

German→French

Ave.	Ave. z	System
77.0	0.249	MSRA-MADL
76.8	0.230	MLLP-UPV
74.8	0.164	Kyoto-University-T2T
75.5	0.160	lingua-custodia-primary
74.4	0.129	LIUM
72.7	0.038	online-B
71.7	0.019	online-Y
68.8	−0.104	TartuNLP-c
66.0	−0.194	online-A
65.0	−0.240	online-G
58.9	−0.456	online-X

French→German

Ave.	Ave. z	System
82.4	0.267	MSRA-MADL
81.5	0.246	eTranslation
78.5	0.082	LIUM
76.8	0.037	MLLP-UPV
76.0	0.001	online-Y
76.6	−0.018	online-G
75.2	−0.034	online-B
74.8	−0.039	online-A
73.9	−0.098	TartuNLP-c
66.5	−0.410	online-X

Table 13: Official results of WMT19 German to French and French to German News Translation Task for which the topic was restricted to EU Elections. 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→English

Ave.	Ave. z	System
75.4	0.283	MSRA-MADL
77.5	0.243	online-B
75.9	0.227	Facebook-FAIR
75.1	0.202	JHU
71.3	0.192	UCAM
77.3	0.171	RWTH-Aachen
76.8	0.166	HUMAN
73.8	0.164	dfki-nmt
77.9	0.162	MLLP-UPV
75.1	0.150	NEU
73.1	0.137	online-Y
72.1	0.103	online-A
71.2	0.009	TartuNLP-c
73.2	−0.052	uedin
67.0	−0.183	online-G
69.0	−0.194	PROMT-NMT
62.8	−0.299	online-X

English→Czech

Ave.	Ave. z	System
84.0	0.915	HUMAN
76.4	0.537	CUNI-Transformer-T2T-2019
76.7	0.528	CUNI-Transformer-T2T-2018
73.7	0.474	CUNI-DocTransformer-T2T
69.7	0.299	CUNI-DocTransformer-Marian
70.0	0.234	uedin
60.0	−0.098	TartuNLP-c
59.9	−0.169	online-Y
57.3	−0.314	online-B
54.7	−0.368	online-G
47.7	−0.619	online-A
47.4	−0.763	online-X

English→German

Ave.	Ave. z	System
82.6	0.530	Facebook-FAIR
81.0	0.335	HUMAN
78.6	0.334	MSRA-MADL
81.3	0.314	Microsoft-WMT19-sent-doc
78.6	0.313	NEU
81.4	0.312	Microsoft-WMT19-doc-level
79.0	0.282	UCAM
77.3	0.268	MLLP-UPV
76.4	0.250	online-Y
78.1	0.200	eTranslation
74.0	0.198	online-B
76.3	0.176	JHU
74.1	0.169	lmu-ctx-tf-single
73.4	0.169	Helsinki-NLP
76.9	0.158	dfki-nmt
76.0	0.156	Microsoft-WMT19-sent-level
73.3	0.101	online-A
73.2	0.058	PROMT-NMT
74.8	0.008	online-G
70.1	−0.027	UdS-DFKI
71.1	−0.087	TartuNLP-c
67.3	−0.285	online-X
40.1	−1.555	en-de-task

English→Finnish

Ave.	Ave. z	System
86.2	1.225	HUMAN
72.9	0.776	GTCOM-Primary
71.0	0.745	MSRA-NAO
57.1	0.293	NICT
57.3	0.237	online-Y
55.1	0.127	Helsinki-NLP
52.2	0.070	online-B
49.6	0.038	online-G
46.2	−0.006	TartuNLP-c
38.0	−0.405	online-A
37.9	−0.433	online-X
39.3	−0.462	Helsinki-NLP-rule-based
14.0	−1.156	apertium-uc

English→Gujarati

Ave.	Ave. z	System
67.1	1.119	HUMAN
57.5	0.759	GTCOM-Primary
63.7	0.737	online-B
54.0	0.561	UEDIN
54.1	0.431	MSRA
47.2	0.146	CUNI-T2T-transfer
44.5	−0.178	Ju-Saarland-clean-num-135-bpe
35.0	−0.481	online-G
33.1	−0.495	IITP-MT
33.0	−0.496	NICT
27.1	−0.724	online-X
29.7	−0.791	UdS-DFKI

English→Kazakh

Ave.	Ave. z	System
73.7	0.883	HUMAN
64.1	0.471	UAlacant-NMT
59.9	0.269	UAlacant-NMT-RBMT
57.9	0.228	MSRA
56.5	0.223	online-B
55.7	0.166	NEU
56.6	0.138	online-G
53.5	0.071	CUNI-T2T-transfer
51.0	−0.039	rug-bpe
45.9	−0.342	talp-upc-2019
37.3	−0.550	NICT
12.2	−1.472	DBMS-KU

English→Lithuanian

Ave.	Ave. z	System
81.2	1.176	HUMAN
63.0	0.548	tilde-nc-nmt
55.4	0.367	MSRA-MASS-uc
58.6	0.342	MSRA-MASS-c
56.9	0.331	tilde-c-nmt
54.6	0.157	GTCOM-Primary
54.3	0.121	eTranslation
51.1	0.040	NEU
48.4	0.017	online-B
39.5	−0.338	TartuNLP-c
28.5	−0.738	online-A
28.8	−0.768	online-X
23.8	−0.797	online-G

English→Russian

Ave.	Ave. z	System
78.9	0.699	HUMAN
78.3	0.645	Facebook-FAIR
72.8	0.449	USTC-MCC
70.8	0.362	online-B
70.8	0.335	online-G
69.4	0.314	NEU
68.0	0.248	PROMT-NMT
65.2	0.157	online-Y
62.7	−0.099	rerank-er
59.9	−0.142	TartuNLP-u
56.8	−0.262	online-A
48.6	−0.389	online-X
32.8	−1.156	NICT

English→Chinese

Ave.	Ave. z	System
70.3	0.486	HUMAN
71.0	0.421	KSAI
69.4	0.303	Baidu
65.6	0.245	NEU
64.7	0.156	BTRANS
65.4	0.146	UEDIN
62.4	0.116	NICT
65.4	0.094	online-A
64.6	0.057	xzl-nmt
59.6	−0.081	online-B
60.5	−0.09	online-Y
58.0	−0.141	online-G
55.3	−0.346	online-X

Chinese→English

Ave.	Ave. z	System
77.7	0.278	Baidu
76.5	0.220	NEU
78.0	0.217	online-B
77.8	0.181	BTRANS-ensemble
74.5	0.169	MSRA-MASS
73.8	0.141	BTRANS
75.6	0.138	KSAI
73.4	0.070	UEDIN
75.6	0.051	online-Y
74.6	0.050	NICT
74.9	0.015	MSRA-MASS
73.4	−0.043	online-A
71.4	−0.104	online-G
67.7	−0.333	online-X
57.8	−0.915	Apprentice-c

Table 14: Document Rating+Document Context (DR+DC) results of WMT19 News Translation Task for subset of language pairs. 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.

Ave.	Ave. z	System
79.1	0.142	NEU
80.9	0.142	KSAI
79.0	0.139	MSRA-MASS
79.5	0.130	online-B
79.5	0.125	Baidu
77.9	0.076	MSRA-MASS
76.0	0.073	BTRANS
77.6	0.051	BTRANS-ensemble
78.0	0.047	online-Y
76.5	−0.015	online-A
75.1	−0.019	UEDIN
75.3	−0.033	NICT
73.3	−0.095	online-G
69.2	−0.276	online-X
58.4	−0.609	Apprentice-c

Table 15: Segment Rating+Document Context (SR+DC) results of WMT19 News Translation Task for Chinese to 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.

In contrast to the previous year, reference translations were scored significantly higher than MT systems in all these settings. It is thus not clear if the super-human quality observed last year was due to lower quality of last year's references, different set of documents or the segment-level style of evaluation as thoroughly discussed by Bojar et al. (2018).

The good news is that all the different types of evaluation correlate very well, with Pearson correlation coefficient ranging from .978 (Ave. of DR+DC vs. SR−DC Microsoft) to .998 (Ave. vs. Ave. z of SR+DC). The document-level ranking (DR+DC) correlates with all variants of segment-level ranking with Pearson of .981 to .996.

4 Test Suites

Following our practice since last year, we issued a call for "test suites", i.e. test sets focussed on particular language phenomena, to complement the standard manual and automatic evaluations of WMT News Translation system.

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 jointly with the regular news texts. 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, we extract translated test suites, unshuffle them and send them back to the corresponding test-suite team. It was up to the test-suite team to evaluate MT outputs and some did this automatically, some manually and some both.

When shuffling, test suites this year closely observed document boundaries. If a test suite was marked as sentence-level only by their authors, we treated individual sentences as if they were one-sentence documents. This lead to a very high number of input documents for some language pairs but all News Task participants managed to handle this additional burden.

As in the previous year, we have to note that test suites go beyond the news domain. If News Task systems are too heavily optimized for news, they may underperform on these domains.

The primary motivation in 2018 was to cut through the opacity of evaluations. We wanted to know more details than just which systems perform better or worse *on average*. This motivation remains also this year but one more reason for people providing test suites was to examine the *human parity* question from additional viewpoints beyond what Bojar et al. (2018) discuss for English→Czech and Hassan et al. (2018) for Chinese→English.

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 Audits and Agreements (Vojtěchová et al., 2019)

The test suite provided by the ELITR project (Vojtěchová et al., 2019) focuses on document-level qualities of two types of documents, audit reports and agreements (represented with only one document, in fact), for the top-performing English-to-Czech systems and some English↔German systems.

The English-to-Czech systems were found as matching or perhaps even surpassing the quality of news reference translations in WMT18 (Bojar et al., 2018) and they also perform very well this year on news. The test suite wanted to validate if this quality transfers (without any specific domain adaptation) also to the domain of reports of supreme audit institutions, which is much more sensitive to terminological choices, and the domain of agreements, where term consistence is

Official SR+DC

Ave.	Ave. z	System
91.2	0.642	HUMAN
86.0	0.402	CUNI-DocTransformer-T2T
86.9	0.401	CUNI-Transformer-T2T-2018
85.4	0.388	CUNI-Transformer-T2T-2019
81.3	0.223	CUNI-DocTransformer-Marian
80.5	0.206	uedin
70.8	−0.156	online-Y
71.4	−0.195	TartuNLP-c
67.8	−0.300	online-G
68.0	−0.336	online-B
60.9	−0.594	online-A
59.3	−0.651	online-X

SR−DC WMT

Ave. z	System
0.62538	HUMAN
0.40505	CUNI-Transformer-T2T-2018
0.39463	CUNI-DocTransformer-T2T
0.35678	CUNI-Transformer-T2T-2019
0.31261	CUNI-DocTransformer-Marian
0.26538	uedin
-0.17006	TartuNLP-c
-0.18841	online-Y
-0.26188	online-B
-0.36871	online-G
-0.67123	online-A
-0.72614	online-X

DR+DC

Ave.	Ave. z	System
84.0	0.915	HUMAN
76.4	0.537	CUNI-Transformer-T2T-2019
76.7	0.528	CUNI-Transformer-T2T-2018
73.7	0.474	CUNI-DocTransformer-T2T
69.7	0.299	CUNI-DocTransformer-Marian
70.0	0.234	uedin
60.0	−0.098	TartuNLP-c
59.9	−0.169	online-Y
57.3	−0.314	online-B
54.7	−0.368	online-G
47.7	−0.619	online-A
47.4	−0.763	online-X

SR−DC Microsoft

Ave. z	System
0.39909	HUMAN
0.30170	CUNI-DocTransformer-T2T
0.28599	CUNI-Transformer-T2T-2018
0.27254	CUNI-Transformer-T2T-2019
0.21186	uedin
0.19160	CUNI-DocTransformer-Marian
-0.05716	TartuNLP-c
-0.09987	online-Y
-0.21633	online-B
-0.29386	online-G
-0.40917	online-A
-0.58836	online-X

Table 16: English-Czech translation: A comparison of SR+DC (official), DR+DC (doclevel) and two versions of segments-evaluation (SR−DC): by WMT annotators and Microsoft annotators.

critical.

The main findings are that also for precise texts (even if intended for the general public and written in a relatively simple language), current NMT systems are close to matching human translation quality. Terminological choices are a little worse but syntax and overall understandability was scored on par or better than the human reference (mixed among the system in an anonymous way). This can be seen as an indication of human parity even out of the original domain of the systems, although the official evaluation on news this year ranks the reference significantly higher.

A very important observation is that (single) reference translations are insufficient because they don't reflect the truly possible term translations. Manual non-expert evaluation would also not be sufficiently reliable because non-experts do not realize the subtle meaning differences among the terms.

On the other hand, the micro-study on agreements reveals that even these very good systems produce practically useless translations of agreements because none of them handles document-specific terms and their consistent translations whatsoever.

4.1.2 Linguistic Evaluation of German-to-English (Avramidis et al., 2019)

The test suite by DFKI covers 107 grammatical phenomena organized into 14 categories. The test suite is very closely related to the one used last year (Macketanz et al., 2018), which allows an evaluation over time.

The test suite is evaluated semi-automatically on a large set of sentences (over 25k) illustrating each of the examined phenomenon and equipped with automatic checks for anticipated good and bad translations. The outputs of these checks are manually verified and refined.

The cross-year comparison is naturally affected by the different set of systems participating in each of the evaluations, but some trends are still observed, namely the improvement in function words, non-verbal agreement and punctuation. The least improvement is seen in terminology and named entities.

Overall, MT system still translate on average about 25% of the tested sentences wrongly. The worst performance is seen for idioms (88% wrong) and complex German verbal grammar (72–77% wrong). Specific terminology and some grammat-

ical phenomena reach about 50%. The paper also indicates phenomena with error rate below 10%, e.g. negation or several cases of verb conjugation.

4.1.3 Document-Level Phenomena (Rysová et al., 2019)

The English-to-Czech test suite by Rysová et al. (2019) builds upon discourse linguistics and manually evaluates three phenomena related to document-level coherence, namely topic-focus articulation (information structure), discourse connectives and alternative lexicalizations of connectives (essentially multi-word discourse connectives). Co-reference is deliberately not included.

The 101 test suite documents (3.5k source sentences in total) come from Penn Discourse Treebank and are speficically the "essay" or "letter" type. The manual evaluation by trained linguists considered always the whole document: the source English text and one of the MT outputs. Targetted phenomena were highlighted in the source and the annotators marked whether they agree with the source annotation and (if yes) whether the respective source phenomenon is also refleted in the target. The reference translation comes from Prague Czech-English Dependency Treebank (Hajič et al., 2012) and it was included in the annotation in a blind way, as if it was one of the MT systems.

The results indicate that the examined phenomena are also handled by the MT systems exceptionally well, matching human quality or even negligibly outperforming humans, e.g. in the mutli-word discourse connectives. Interestingly, the English-Czech systems trained in some document-level way this year do not seem any better than the segment-level ones.

4.1.4 Producing German Conjuctions from English and French (Popović, 2019)

The test suite by Popović (2019) contains approximately 1000 English and 1000 individual French sentences that were included in the English→German and French→German tasks. The sentences focus on the translation of the English "but" and French "mais" which should be disambiguated into German "aber" or "sondern".

Except for 1–2% of cases (when no conjunction or both possibilities are found in the target), the outputs can be evaluated automatically. The results indicate that the situation when "aber" is needed is recognized almost perfectly by all the system but the situation which requires "sondern" is sometimes mishandled and the (generally more frequent) "aber" is used. The error rate ranges from 3% (TARTUNLP-C) to 14% (ONLINE-X) or 22% (the unclear system called EN-DE-TASK)

4.1.5 Out-of-Domain Check of Formal Language for German→English (Biçici, 2019)

A small test suite by Biçici (2019) contains 38 sentences from texts by Prussian Cultural Heritage Foundation, checking the performance of MT systems on the domain of cross-cultural international relations.

The test suite is evaluated only with a few automatic measures with no clear conclusion.

4.1.6 Word Sense Disambiguation (Raganato et al., 2019)

Raganato et al. (2019) present the MuCoW (multilingual contrastive word sense disambiguation) test suite which contains a relatively large set of sentences (69–4268 depending on the language pair) mined from parallel corpora to illustrate words which are particularly ambiguous for the given translation pair.

Originally, the test suite relies on MT systems scoring candidate pairs of sentences. Raganato et al. (2019) adapt it for the use case of WMT test suites where the black-box MT systems only provide their translation output. Due care is taken in sentence selection, in particular any overlap with WMT constrained training data is avoided.

The test suite covers from German, Finnish, Lithuanian and Russian into English and from English into these four langauges and Czech.

The ambiguous words were identified with the help of BabelNet (Navigli and Ponzetto, 2012) multilingual synsets and the granularity was reduced with the help of word embeddings to ensure that the meaning distinctions are reliably big. For the WMT use case, there are dozens or a few hundreds of ambiguous source words (except Lithuanian with only very few words) with slightly more than 2 distinct word senses per examined source word on average.

The results show that overall, WMT systems perform quite well word-sense disambiguation when evaluated in the "in-domain" setting (word senses not too common in subtitle corpora), with precision (examples with correct target words over examples with either correct or in-

correct target words) in the ranges 64–80% (e.g. Finnish→English or English→German) up to 95–97% (English→Czech) depending on the language pair. The recalls (examples with correct target words over all examples) are similarly high, 65–91 across the board.

The "out-of-domain" evaluation was directed at word senses common in colloquial speech and in general, research WMT news system perform a little worse than online systems in these scores except for English-Czech.

5 Similar Language Translation

Within the MT and NLP communities, English is by far the most resource-rich language. MT systems are most often trained to translate texts from and to English or they use English as a pivot language to translate between resource-poorer languages. The interest in English is reflected, for example, in the WMT translation tasks (e.g. News, Biomedical) which have always included language pairs in which texts are translated to and/or from English.

With the widespread use of MT technology, there is more and more interest in training systems to translate between languages other than English. One evidence of this is the need of directly translating between pairs of similar languages, varieties, and dialects (Zhang, 1998; Marujo et al., 2011; Hassani, 2017; Costa-jussà et al., 2018). The main challenge is to take advantage of the similarity between languages to overcome the limitation given the low amount of available parallel data to produce an accurate output.

Given the interest of the community in this topic we organize, for the first time at WMT, a shared task on "Similar Language Translation" to evaluate the performance of state-of-the-art translation systems on translating between pairs of languages from the same language family. We provide participants with training and testing data from three language pairs: Spanish - Portuguese (Romance languages), Czech - Polish (Slavic languages), and Hindi - Nepali (Indo-Aryan languages). Evaluation will be carried out using automatic evaluation metrics and human evaluation.

5.1 Data

Training We have made available a number of data sources for the Similar Language Translation shared task. Some training datasets were used in the previous editions of the WMT News Translation shared task and were updated (Europarl v9, News Commentary v14), while some corpora were newly introduced (Wiki Titles v1, JRC Acquis). For the Hi–Ne language pair, parallel corpora have been collected from Opus (Tiedemann and Nygaard, 2004)[19]. We used the Ubuntu, KDE, and Gnome corpus available at OPUS for this shared task.

Development and Test Data The creation of development and test sets for Czech and Polish involved random extraction of 30 TED talks for the development and 30 TED talks for the test set in each language. Then unique sentences were extracted and cleaning of lines containing meta-data information was performed which resulted in 4.7k sentences in the development sets and 4.8k sentences in the test sets. Further cleaning of the corpus to retain only sentences between 7 and 100 words limited the number of the sentences in the dev and test sets to 3050 and 3412 sentences respectively.

The development and test sets for Spanish and Portuguese were created from a corpus provided by AT Language Solutions [20]. First, the extraction of unique sentences and cleaning of lines containing meta-data information was performed which narrowed the number of sentences to 11.7k sentences. Then cleaning of the corpus to retain only sentences between 7 and 100 words limited the number of the sentences to 6.8k. Finally, 3k randomly selected sentences were used for the development set and other 3k random sentences were extracted to form the test set. For HI–NE, all data was initially combined and randomly shuffled. From the combined corpus, we randomly extracted 65,505 sentences for the training set, 3,000 sentences for development set and 3,567 for the test set. Finally, the test set was split into two different test sets: 2,000 sentences used for HI to NE and 1,557 sentences were used for NE to HI.

5.2 Participants

The first edition of the WMT Similar Language Translation task attracted more participants than we anticipated. There were 35 teams who signed up to participate in the competition and 14 of them submitted their system outputs to one of the three language pairs in any translation direction. In the

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

[20] https://www.at-languagesolutions.com/en/

Table 17: Europarl v9 Parallel Corpus

	Czech ↔ Polish		Spanish ↔ Portuguese	
sentences	631372		1811977	
words	12526659	12641841	47832025	46191472

Table 18: Wiki Titles v1 Parallel Corpus

	Czech ↔ Polish		Spanish ↔ Portuguese	
sentences	248645		621296	
words	551084	554335	1564668	1533764

Table 19: JRC-Acquis Parallel Corpus

	Czech ↔ Polish		Spanish ↔ Portuguese	
sentences	1311362		1650126	
words	21409363	21880482	35868080	33474269

Table 20: News Commentary v14 Parallel Corpus

	Spanish ↔ Portuguese	
sentences	48168	
words	1271324	1219031

Table 21: GNOME, Ubuntu, KDE Parallel Corpus

	Hindi ↔ Nepali	
sentences	65505	
words	253216	222823

Table 22: Europarl v9 Monolingual Corpus

	Czech	Polish	Spanish	Portuguese
sentences	665433	382726	2019336	2015290
words	13199347	7087267	52157546	50462045

Table 23: News Crawl Monolingual Corpus

	Czech	Polish	Spanish	Portuguese
sentences	72157988	814754	43814290	8301536
words	1019497060	12370354	1159300825	160477593

Table 24: News Commentary v14 Monolingual Corpus

	Czech	Spanish	Portuguese
sentences	266705	424063	59502
words	4922572	10724738	1443204

end of the competition, 10 teams submitted system description papers which are referred to in this report. Table 25 summarizes the participation across language pairs and translation directions and includes references to the 10 system description papers.

We observed that the majority of teams contain only members which work in universities and research centers (12 teams) whereas only two teams contain members who work in the industry. The participants were distributed across different continents with a higher participation of European teams (7 European) with two teams based on the Americas, and five Asian teams.

As follows we provide summaries for each of the entries we received:

BSC: Team BSC (Barcelona SuperComputing Center) participated with a Transfomer-based approach in the Spanish-Portuguese track. As preprocessing, SentencePiece [21] was applied after concatenating and shuffling the data. For the Portuguese to Spanish language direction, BSC made use of back-translation.

CFILT_IITB: The CFILT_IITB submission (Khatri and Bhattacharyya, 2019) is based on unsupervised neural machine translation described in Artetxe et al. (2018) in the task Hindi ↔ Nepali, where encoder is shared and following bidirectional recurrent neural network architecture. They used 2 hidden layers for both encoder and decoder.

CMUMEAN: The is system is based on standard

[21] https://github.com/google/sentencepiece

transformer based NMT model for the Hindi ↔ Nepali shared task. To compensate the insufficient released parallel data, they utilized 7M monolingual data for both Hindi and Nepali taken from CommonCrawl. They augmented the monolingual data by constructing pseudo-parallel datasets. The pseudo-parallel sentences were constructed by word substitutions, based on a mapping of the embedding spaces of the two languages. These mapping were learned from all data and a seed dictionary based on the alignment of the parallel data.

Incomslav: Team INCOMSLAV (Chen and Avgustinova, 2019) by Saarlad University participated in the Czech to Polish translation task only. The team's primary submission builds on a transformer-based NMT baseline with back translation which has been submitted one of their contrastive submission. Incomslav's primary system is a phoneme-based system re-scored using their NMT baseline. A second contrastive submission builds our phrase-based SMT system combined with a joint BPE model.

JUMT: This submittion used phrase based statistical machine translation model for Hindi → Nepali task. They used 3-gram language model and MGIZA++ for word alignment. However, their system achieved poor performance in the shared task.

MLLP-UPV: Team MLLP-UPV (Baquero-Arnal et al., 2019) by Universitat Politècnica de Valèncía (UPV) participated with a Transformer (implemented with FairSeq (Ott et al., 2019)) and a fine-tuning strategy for domain adaptaion in the task of Spanish-Portuguese. Fine-tunning on the development data provide improvements of almost 12 BLEU points, which may explain their clear best performance in the task for this language pair. As a contrastive system authors provided only for the Portuguese-to-Spanish a novel 2D alternating RNN model which did not respond so well when fine-tunning.

KYOTOUNIVERSITY: Kyoto University's submission, listed simply as KYOTO in Table 25 for PT → ES task is based on transformer NMT system. They used difference word segmentation strategies during preprocessing. Additionally they used optional reverse feature in their preprocessing step. Their submission achieved average scores in the shared task.

NICT: The NICT team (Marie et al., 2019a) participated with the a system combination between the Transformer (implemented in Marian (Junczys-Dowmunt et al., 2018) and Phrase-based machine translation system (implemented with Moses) and for the Spanish-Portuguese task. The system combination included features formerly presented in (Marie and Fujita, 2018), including scores left-to-right and right-to-left, sentence level translation probabilities and language model scores. Also authors provide contrastive results with an unsupervised phrase-based MT system which achieves quite close results to their primary system. Authors associate high performance of the unsupervised system to the language similarity.

NITS-CNLP: The NITS-CNLP team (Laskar et al., 2019) by the National Institute of Technology Silchar in India submitted results to the HI-NE translation task in both directions. The NITS-CNLP systems are based on Marian NMT (Junczys-Dowmunt et al., 2018) and Open NMT implementations of sequence-to-sequence RNNs with attention mechanisms. Their contrastive submissions were ranking first in both Hindi to Nepali and Nepali to Hindi translation.

Panlingua-KMI: The Panlingua-KMI team (Ojha et al., 2019) tested phrase-based SMT and NMT methods for HI-NE translation in both directions. The PBSMT systems have been trained using Moses (Koehn et al., 2007) and KenLM. Their two NMT systems were built using OpenNMT. The first system was built with 2 layers using LSTM model while the second system was built with 6 layers using the Transformer model.

UBC-NLP: Team UBC-NLP from the University of British Columbia in Canada (Przystupa and Abdul-Mageed, 2019) compared the performance of the LSTM plus attention (Bahdanau et al., 2015) and Transformer (Vaswani et al., 2017) (implemented in OpenNMT toolkit[22]) perform for the three tasks at hand. Authors use backtranslation to introduce monolingual data in their systems. LSTM plus attention outperformed Transformer for Hindi-Nepali, and viceversa for the other two tasks. As reported by the authors, Hindi-Nepali task provides much more shorter sentences than

[22] http://opennmt.net/

the other two-tasks. Additionally, authors in their system description report interesting insights on how similar are languages in each of the 3 different tasks.

UDS-DFKI: The UDS-DFKI team (Pal et al., 2019) is formed by researchers from Saarland University (UDS), the German Research Foundation of Artificial Intelligence (DFKI), and the University of Wolverhampton. They submitted a *transference* model that extends the original transformer model to multi-encoder based transformer architecture. The *transference* model contains two encodes, the first encoder encodes word form information of the source (CS), and a second encoder to encode sub-word (byte-pair-encoding) information of the source (CS). The results obtained by their system in translating from Czech→Polish and comment on the impact of out-of-domain test data in the performance of their system. UDS-DFKI ranked second among ten teams in Czech–Polish translation.

UHelsinki: The University of Helsinki team (Scherrer et al., 2019) participated with the Transformer (Vaswani et al., 2017) implemented in the OpenNMT toolkit. They focused on word segmentation methods and compared a cognate-aware segmentation method, Cognate Morfessor (Grönroos et al., 2018), with character segmentation and unsupervised segmentation methods. As primary submission they submitted this Cognate Morfessor that optimizes subword segmentations consistently for cognates. They participated for all translation directions in Spanish-Portuguese and Czech-Polish, and this Cognate Morfessor performed better for Czech-Polish, while character-based segmentations (Costa-jussà and Fonollosa, 2016), while much more inefficient, were superior for Spanish-Portuguese.

UPC-TALP: The UPC-TALP team (Biesialska et al., 2019) by the Universitat Politècnica de Catalunya submitted a Transformer (implemented with Fairseq (Ott et al., 2019)) for the Czech-to-Polish task and a Phrase-based system (implemented with Moses (Koehn et al., 2007)) for Spanish-to-Portuguese. They tested adding monolingual data to the NMT system by copying the same data on the source and target sides, with negative results. Also, their system combination based on sentence-level BLEU in back-translation did not succeed. Authors provide interesting insights on language distance based on previous work by (Gamallo et al., 2017) and their results show that the Phrase-based compared to NMT achieves better results when the language distance between source and target language is lower.

5.3 Results

We present results for the three language pairs, each of them in the two possible directions. For this first edition of the Similar Translation Task and differently from News task, evaluation was only performed on automatic basis using BLEU (Papineni et al., 2002) 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 (C) or primary (P), and the BLEU and TER results. In general, primary systems tend to be better than contrastive systems, as expected, but there are some exceptions.

Even if we are presenting 3 pairs of languages each pair belonging to the same family, translation quality in terms of BLEU varies signficantly. While the best systems for Spanish-Portuguese are above 64 BLEU and below 21 TER (see Tables 26 and 27), best systems for Czech-Polish do not reach the 8 BLEU and the 79.6 TER for the direction with lowest TER (Polish-to-Czech). The case of Hindi-Nepali is in between, with BLEU of 53.7 and TER of 36.3 for the better direcion Hindi-to-Nepali. Also, we noticed that BLEU and TER do not always correlate and while some systems performed better in BLEU, the ranking is different if ordered by TER. In any case, we chose BLEU as the official metric for ranking.

The highest variance of system performance can be found in Hindi-Nepali (both directions), where the best performing system is around 50 BLEU (53 for Hindi-to-Nepali and 49.1 for Nepali-to-Hindi), and the lowest entry is 1.4 for Hindi-to-Nepali and 0 for Nepali-to-Hindi. The lowest variance is for Polish-to-Czech and it may be because only two teams participated.

5.4 Conclusion of Similar Language Translation

In this section we presented the results of the WMT Similar Language Translation shared task 2019. The competition featured data in three language pairs: Czech-Polish, and Hindi-Nepali, and Portuguese-Spanish.

Team	CS→PL	PL→CS	HI→NE	NE→HI	PT→ES	ES→PT	Paper
BSC					✓	✓	
CFILT_IITB			✓	✓			Khatri and Bhattacharyya (2019)
CMUMEAN			✓	✓			
Incomslav	✓						Chen and Avgustinova (2019)
JUMT			✓				
KYOTO					✓		
MLLP-UPV					✓	✓	Baquero-Arnal et al. (2019)
NICT					✓		Marie et al. (2019a)
NITS-CNLP			✓	✓			Laskar et al. (2019)
Panlingua-KMI			✓	✓			Ojha et al. (2019)
UBC-NLP	✓	✓	✓	✓	✓	✓	Przystupa and Abdul-Mageed (2019)
UDS-DFKI	✓						Pal et al. (2019)
UHelsinki	✓	✓			✓	✓	Scherrer et al. (2019)
UPC-TALP	✓					✓	Biesialska et al. (2019)
Total	5	2	6	5	6	5	10

Table 25: The teams that participated in the Similar Translation Task.

Team	Type	BLEU	TER
MLLPUPV	P	66.6	19.7
NICT	P	59.9	25.3
Uhelsinki	C	59.1	25.5
Uhelsinki	C	58.6	25.1
Uhelsinki	P	58.4	25.3
KYOTOUNIVERSITY	P	56.9	26.9
NICT	C	54.9	28.4
BSC	P	54.8	29.8
UBC-NLP	P	52.3	32.9
UBC-NLP	C	52.2	32.8
MLLPUPV	C	51.9	30.5
MLLPUPV	C	49.7	32.1
BSC	C	48.5	35.1

Table 26: Results for Portuguese to Spanish Translation

Team	Type	BLEU	TER
MLLPUPV	P	64.7	20.8
UPC-TALP	P	62.1	23.0
NICT	P	53.3	29.1
Uhelsinki	C	52.8	28.6
Uhelsinki	P	52.0	29.4
Uhelsinki	C	51.0	33.1
NICT	C	47.9	33.4
UBC-NLP	P	46.1	36.0
UBC-NLP	C	46.1	35.9
MLLPUPV	C	45.5	35.3
BSC	P	44.0	37.5

Table 27: Results for Spanish to Portuguese Translation

Team	Type	BLEU	TER
NITS-CNLP	C	53.7	36.3
Panlingua-KMI	P	11.5	79.1
CMUMEAN	P	11.1	79.7
UBC-NLP	P	08.2	77.1
UBC-NLP	C	08.2	77.2
NITS-CNLP	P	03.7	-
NITS-CNLP	C	03.6	-
CFILT_IITB	C	03.5	-
Panlingua-KMI	C	03.1	-
CFILT_IITB	P	02.8	-
CFILT_IITB	C	02.7	-
Panlingua-KMI	C	01.6	-
JUMT	P	01.4	-

Table 28: Results for Hindi to Nepali Translation

Team	Type	BLEU	TER
NITS-CNLP	C	49.1	43.0
NITS-CNLP	P	24.6	69.1
CMUMEAN	P	12.1	76.2
Panlingua-KMI	P	09.8	91.3
UBC-NLP	P	09.1	88.3
UBC-NLP	C	09.1	88.4
Panlingua-KMI	C	04.2	-
Panlingua-KMI	C	03.6	-
CFILT_IITB	P	02.7	-
NITS-CNLP	C	01.4	-
CFILT_IITB	C	0	-
CFILT_IITB	C	0	-

Table 29: Results for Nepali to Hindi Translation

For the future it is worth investigating why languages from the same family, like Czech-Polish have extremely low performance. Authors in (Biesialska et al., 2019), with the best performing system in Czech-to-Polish, hypothesize that one of the reasons is the different in alphabets from both languages. Additionally, they refer to

Team	Type	BLEU	TER
UPC-TALP	P	7.9	85.9
UDS-DFKI	P	7.6	87.0
Uhelsinki	P	7.1	87.4
Uhelsinki	C	7.0	87.3
Incomslav	C	5.9	88.4
Uhelsinki	C	5.9	88.4
Incomslav	P	3.2	-
Incomslav	C	3.1	-
UBC-NLP	C	2.3	-
UBC-NLP	P	2.2	-

Table 30: Results for Czech to Polish Translation

Team	Type	BLEU	TER
Uhelsinki	C	7.2	79.6
Uhelsinki	P	7.0	79.4
UBC-NLP	P	6.9	86.5
UBC-NLP	C	6.9	86.2
Uhelsinki	C	6.6	80.2

Table 31: Results for Polish to Czech Translation

Gamallo et al. (2017) and provide big language distances for Czech-Polish compared to Spanish-Portuguese.

6 Conclusion

We presented the results of the WMT18 News Translation Shared Task. Our main findings rank participating systems in their sentence-level translation quality, as assessed in a large-scale manual evaluation using the method of Direct Assessment (DA).

The novelties this year include (1) avoiding effects of translationese by creating reference translations always in the same directions as the MT systems are run, (2) providing human assessors with the context of the whole document when assessing individual segments for a large portion of language pairs, (3) extending the set of languages which are evaluated given the *source*, not the reference translation, and (4) scoring also whole documents, not only individual segments.

Our results indicate which MT systems perform best across the 18 examined translation pairs, as well as what features are now commonly used in the field. The test suites complement this evaluation by focussing on particular language phenomena such as word-sense disambiguation, document-level coherence or terminological correctness.

As in the previous year, MT systems seem to reach the quality of human translation in the news domain for some language pairs. This result has to be regarded with a great caution and considering the technical details of the (document-aware) DA evaluation method as well as the outcomes of complementary evaluations, such as those included in the test suites. Importantly, the language pairs where the parity was reached last year were not confirmed by the evaluation this year and a similar situation can repeat. As one of the test suites (Vojtěchová et al., 2019) suggests, there are aspects of texts which are wrongly handled by even the best translation systems.

The task on similar language translation indicated that the performance in this area is extremely varied across language pairs as well as across participating teams.

Acknowledgments

This work was supported in part by funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nos. 825299 (GoURMET) and 825303 (Bergamot), and from the Connecting Europe Facility under agreement No. NEA/CEF/ICT/A2016/1331648 (ParaCrawl).

The human evaluation campaign was very gratefully supported by Apple, Microsoft, and Science Foundation Ireland in the ADAPT Centre for Digital Content Technology (www.adaptcentre.ie) at Dublin City University funded under the SFI Research Centres Programme (Grant 13/RC/2106) co-funded under the European Regional Development Fund.

We are grateful to the large number of anonymous Mechanical Turk workers who contributed their human intelligence to the human evaluation.

Ondřej Bojar would like to acknowledge also the grant no. 19-26934X (NEUREM3) of the Czech Science Foundation.

The organizers of the similar languages task want to thank Magdalena Biesialska for her support in the compilation of the Czech-Polish data set as well as her valuable support as a Polish native speaker. This work is supported in part by the Spanish Ministerio de Economía y Competitividad, the European Regional Development Fund and the Agencia Estatal de Investigación, through the postdoctoral senior grant Ramón y Cajal, through the travel grant José Castillejo, CAS18/00223, the contract TEC2015-69266-P

(MINECO/FEDER,EU) and the contract PCIN-2017-079 (AEI/MINECO). The authors also want to thank German research foundation (DFG) under grant number GE 2819/2-1 (project MMPE) and the German Federal Ministry of Education and Research (BMBF) under funding code 01IW17001 (project Deeplee).

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *Proceedings of the Sixth International Conference on Learning Representations*.

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.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Pau Baquero-Arnal, Javier Iranzo-Sánchez, Jorge Civera, and Alfons Juan. 2019. The MLLP-UPV Spanish-Portuguese and Portuguese-Spanish Machine Translation Systems for WMT19 Similar Language Translation Task. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Rachel Bawden, Nikolay Bogoychev, Ulrich Germann, Roman Grundkiewicz, Faheem Kirefu, Antonio Valerio Miceli Barone, and Alexandra Birch. 2019a. The University of Edinburgh's Submissions to the WMT19 News Translation Task. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurelie Neveol, Mariana Neves, Felipe Soares, Amy Siu, Karin Verspoor, and Maika Vicente Navarro. 2019b. Findings of the WMT 2019 Biomedical Translation Shared Task: Evaluation for MEDLINE Abstracts and Biomedical Terminologies. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Chao Bei, Hao Zong, Conghu Yuan, Qingming Liu, and Baoyong Fan. 2019. GTCOM Neural Machine Translation Systems for WMT19. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Magdalena Biesialska, Lluis Guardia, and Marta R. Costa-jussà. 2019. The TALP-UPC System for the WMT Similar Language Task: Statistical vs Neural Machine Translation. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Ergun Biçici. 2019. Machine Translation with parfda, Moses, kenlm, nplm, and PRO. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Nikolay Bogoychev, Marcin Junczys-Dowmunt, Kenneth Heafield, and Alham Fikri Aji. 2018. Accelerating asynchronous stochastic gradient descent for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium.

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.

Fethi Bougares, Jane Wottawa, Anne Baillot, Loïc Barrault, and Adrien Bardet. 2019. LIUM's Contributions to the WMT2019 News Translation Task: Data and Systems for German-French Language Pairs. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Eleftheria Briakou and Marine Carpuat. 2019. The University of Maryland's Kazakh-English Neural Machine Translation System at WMT19. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Sari Dewi Budiwati, Al Hafiz Akbar Maulana Siagian, Tirana Noor Fatyanosa, and Masayoshi Aritsugi. 2019. DBMS-KU Interpolation for WMT19 News Translation Task. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Franck Burlot. 2019. Lingua Custodia at WMT'19: Attempts to Control Terminology. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

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

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

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

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–48, Montreal, Canada. Association for Computational Linguistics.

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

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

Noe Casas, José A. R. Fonollosa, Carlos Escolano, Christine Basta, and Marta R. Costa-jussà. 2019. The TALP-UPC Machine Translation Systems for WMT19 News Translation Task: Pivoting Techniques for Low Resource MT. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Rajen Chatterjee, Christian Federmann, Matteo Negri, and Marco Turchi. 2019. Findings of the WMT 2019 Shared Task on Automatic Post-Editing. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Yu Chen and Tania Avgustinova. 2019. Machine Translation from an Intercomprehension Perspective. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Marta R. Costa-jussà and José A. R. Fonollosa. 2016. Character-based neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 357–361, Berlin, Germany. Association for Computational Linguistics.

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

Fabien Cromieres and Sadao Kurohashi. 2019. Kyoto University Participation to the WMT 2019 News

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

Raj Dabre, Kehai Chen, Benjamin Marie, Rui Wang, Atsushi Fujita, Masao Utiyama, and Eiichiro Sumita. 2019. NICT's Supervised Neural Machine Translation Systems for the WMT19 News Translation Task. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Liang Ding and Dacheng Tao. 2019. The University of Sydney's Machine Translation System for WMT19. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding Back-Translation at Scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Brussels, Belgium. Association for Computational Linguistics.

Cristina España-Bonet and Dana Ruiter. 2019. UdS-DFKI Participation at WMT 2019: Low-Resource (en-gu) and Coreference-Aware (en-de) Systems. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. 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.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 Shared Task on Quality Estimation. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Pablo Gamallo, José Ramom Pichel, and Iñaki Alegria. 2017. From language identification to language distance. *Physica A: Statistical Mechanics and its Applications*, 484:152 – 162.

Vikrant Goyal and Dipti Misra Sharma. 2019. The IIIT-H Gujarati-English Machine Translation System for WMT19. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous Measurement Scales in Human Evaluation of Machine Translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Yvette Graham, 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. 2019a. Translationese in machine translation evaluation. *CoRR*, abs/1906.09833.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2019b. Translationese in machine translation evaluation and mt checklist. *CoRR*.

Stig-Arne Grönroos, Sami Virpioja, and Mikko Kurimo. 2018. Cognate-aware morphological segmentation for multilingual neural translation. *CoRR*, abs/1808.10791.

Xinze Guo, Chang Liu, Xiaolong Li, Yiran Wang, Guoliang Li, Feng Wang, Zhitao Xu, Liuyi Yang, Li Ma, and Changliang Li. 2019. Kingsoft's Neural Machine Translation System for WMT19. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Jeremy Gwinnup, Grant Erdmann, and Tim Anderson. 2019. The AFRL WMT19 Systems: Old Favorites and New Tricks. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, and Zdeněk Žabokrtský. 2012. Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the Eighth International Language Resources and Evaluation Conference (LREC'12)*, pages 3153–3160, Istanbul, Turkey. ELRA, European Language Resources Association.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. https://www.microsoft.com/en-us/research/uploads/prod/2018/03/final-achieving-human.pdf.

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

Chris Hokamp, John Glover, and Demian Gholipour Ghalandari. 2019. Evaluating the Supervised and Zero-shot Performance of Multilingual Translation Models. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Javier Iranzo-Sánchez, Gonçal Garcés Díaz-Munío, Jorge Civera, and Alfons Juan. 2019. The MLLP-UPV Supervised Machine Translation Systems for WMT19 News Translation Task. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2018a. Dual Conditional Cross-Entropy Filtering of Noisy Parallel Corpora. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2018b. Microsoft's Submission to the WMT2018 News Translation Task: How I Learned to Stop Worrying and Love the Data. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2019. Microsoft Translator at WMT 2019: Towards Large-Scale Document-Level Neural Machine Translation. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast Neural Machine Translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121. Association for Computational Linguistics.

Jyotsana Khatri and Pushpak Bhattacharyya. 2019. Utilizing Monolingual Data in NMT for Similar Languages: Submission to Similar Language Translation Task. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 114 13:3521–3526.

Tom Kocmi and Ondřej Bojar. 2019. CUNI Submission for Low-Resource Languages in WMT News 2019. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Philipp Koehn, Kevin Duh, and Brian Thompson. 2018. The JHU Machine Translation Systems for WMT 2018. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino. 2019. Findings of the WMT 2019 Shared Task on Parallel Corpus Filtering for Low-Resource Conditions. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. 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.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1318–1326, Stroudsburg, PA, USA. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *CoRR*, abs/1808.06226.

Ivana Kvapilíková, Dominik Macháček, and Ondřej Bojar. 2019. CUNI Systems for the Unsupervised News Translation Task in WMT 2019. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Sahinur Rahman Laskar, Partha Pakray, and Sivaji Bandyopadhyay. 2019. Neural Machine Translation: Hindi-Nepali. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

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.

Bei Li, Yinqiao Li, Chen Xu, Ye Lin, Jiqiang Liu, Hui Liu, Ziyang Wang, Yuhao Zhang, Nuo Xu, Zeyang Wang, Kai Feng, Hexuan Chen, Tengbo Liu, Yanyang Li, Qiang Wang, Tong Xiao, and Jingbo Zhu. 2019a. The NiuTrans Machine Translation Systems for WMT19. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Xian Li, Paul Michel, Antonios Anastasopoulos, Yonatan Belinkov, Nadir Durrani, Orhan Firat, Philipp Koehn, Graham Neubig, Juan Pino, and Hassan Sajjad. 2019b. Findings of the First Shared Task on Machine Translation Robustness. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Zhenhao Li and Lucia Specia. 2019. A Comparison on Fine-grained Pre-trained Embeddings for the WMT19Chinese-English News Translation Task. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Patrick Littell, Chi-kiu Lo, Samuel Larkin, and Darlene Stewart. 2019. Multi-Source Transformer for Kazakh-Russian-English Neural Machine Translation. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Zihan Liu, Yan Xu, Genta Indra Winata, and Pascale Fung. 2019. Incorporating Word and Subword Units in Unsupervised Machine Translation Using Language Model Rescoring. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Renqian Luo, Fei Tian, Tao Qin, Enhong Chen, and Tie-Yan Liu. 2018. Neural architecture optimization. In *Advances in neural information processing systems*, pages 7816–7827.

Jerry Ma and Denis Yarats. 2018. Quasi-hyperbolic momentum and Adam for deep learning. *arXiv preprint arXiv:1810.06801*.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT 2019 Metrics Shared Task: Segment-Level and Strong MT Systems Pose Big Challenges. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. 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.

Kelly Marchisio, Yash Kumar Lal, and Philipp Koehn. 2019. Johns Hopkins University Submission for WMT News Translation Task. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Benjamin Marie, Raj Dabre, and Atsushi Fujita. 2019a. NICT's Machine Translation Systems for the WMT19 Similar Language Translation Task. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Benjamin Marie and Atsushi Fujita. 2018. A smorgasbord of features to combine phrase-based and neural machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, page 111–124, Boston, MA. Association for Machine Translation in the Americas.

Benjamin Marie, Haipeng Sun, Rui Wang, Kehai Chen, Atsushi Fujita, Masao Utiyama, and Eiichiro Sumita. 2019b. NICT's Unsupervised Neural and Statistical Machine Translation Systems for the WMT19 News Translation Task. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. 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*.

Alexander Molchanov. 2019. PROMT Systems for WMT 2019 Shared Translation Task. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Riktim Mondal, Shankha Raj Nayek, Aditya Chowdhury, Santanu Pal, Sudip Kumar Naskar, and Josef van Genabith. 2019. JU-Saarland Submission to the WMT2019 English–Gujarati Translation Shared Task. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 News Translation Task Submission. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Atul Kr. Ojha, Ritesh Kumar, Akanksha Bansal, and Priya Rani. 2019. Panlingua-KMI MT System for Similar Language Translation Task at WMT 2019. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Csaba Oravecz, Katina Bontcheva, Adrien Lardilleux, László Tihanyi, and Andreas Eisele. 2019. eTranslation's Submissions to the WMT 2019 News Translation Task. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Santanu Pal, Marcos Zampieri, and Josef van Genabith. 2019. UDS–DFKI Submission to the WMT2019 Czech–Polish Similar Language Translation Shared Task. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Philadelphia, Pennsylvania.

Mārcis Pinnis. 2018. Tilde's Parallel Corpus Filtering Methods for WMT 2018. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Mārcis Pinnis, Rihards Krišlauks, Daiga Deksne, and Toms Miks. 2017. Neural Machine Translation for Morphologically Rich Languages with Improved Sub-word Units and Synthetic Data. In *Proceedings of the 20th International Conference of Text, Speech and Dialogue (TSD2017)*, volume 10415 LNAI, Prague, Czechia.

Marcis Pinnis, Rihards Krišlauks, and Matiss Rikters. 2019. Tilde's Machine Translation Systems for WMT 2019. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Mārcis Pinnis, Andrejs Vasiljevs, Rihards Kalniņš, Roberts Rozis, Raivis Skadiņš, and Valters Šics. 2018. Tilde MT Platform for Developing Client Specific MT Solutions. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Tommi Pirinen. 2019. Apertium-fin-eng–Rule-based Shallow Machine Translation for WMT 2019 Shared Task. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Martin Popel. 2018. CUNI Transformer Neural MT System for WMT18. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Martin Popel, Dominik Macháček, Michal Auersperger, Ondřej Bojar, and Pavel Pecina. 2019. English-Czech Systems in WMT19: Document-Level Transformer. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Maja Popović. 2019. Evaluating Conjunction Disambiguation on English-to-German and French-to-German WMT 2019 Translation Hypotheses. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Michael Przystupa and Muhammad Abdul-Mageed. 2019. Neural Machine Translation of Low-Resource and Similar Languages with Backtranslation. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2019. The MuCoW Test Suite at WMT 2019: Automatically Harvested Multilingual Contrastive Word Sense Disambiguation Test Sets for Machine Translation. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

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

Jan Rosendahl, Christian Herold, Yunsu Kim, Miguel Graça, Weiyue Wang, Parnia Bahar, Yingbo Gao, and Hermann Ney. 2019. The RWTH Aachen University Machine Translation Systems for WMT 2019. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Kateřina Rysová, Lucie Rysová, Magdaléna Poláková, Tomáš Musil, and Ondřej Bojar. 2019. Manual Evaluation of Discourse Relations Translation Accurracy in Document Level NMT. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Yves Scherrer, Raúl Vázquez, and Sami Virpioja. 2019. The University of Helsinki Submissions to

the WMT19 Similar Language Translation Task. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Sukanta Sen, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2019. IITP-MT System for Gujarati-English News Translation Task at WMT 2019. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017. The University of Edinburgh's Neural MT Systems for WMT17. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Leslie N. Smith. 2018. A disciplined approach to neural network hyper-parameters: Part 1 - learning rate, batch size, momentum, and weight decay. *CoRR*, abs/1803.09820.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, pages 223–231.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pages 5926–5936.

Felix Stahlberg, Adrià de Gispert, Eva Hasler, and Bill Byrne. 2017. Neural machine translation by minimising the Bayes-risk with respect to syntactic translation lattices. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 362–368, Valencia, Spain. Association for Computational Linguistics.

Felix Stahlberg, Danielle Saunders, Adrià de Gispert, and Bill Byrne. 2019. CUED@WMT19:EWC&LMs. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Dario Stojanovski and Alexander Fraser. 2019. Combining Local and Document-Level Context: The LMU Munich Neural Machine Translation System at WMT19. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Dario Stojanovski, Viktor Hangya, Matthias Huck, and Alexander Fraser. 2019. The LMU Munich Unsupervised Machine Translation System for WMT19. In *Proceedings of the Fourth Conference on Machine Translation*, 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.

Meng Sun, Bojian Jiang, Hao Xiong, Zhongjun He, Hua Wu, and Haifeng Wang. 2019. Baidu Neural Machine Translation Systems for WMT19. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Víctor M. Sánchez-Cartagena, Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. 2019. The Universitat d'Alacant Submissions to the English-to-Kazakh News Translation Task at WMT 2019. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Aarne Talman, Umut Sulubacak, Raúl Vázquez, Yves Scherrer, Sami Virpioja, Alessandro Raganato, Arvi Hurskainen, and Jörg Tiedemann. 2019. The University of Helsinki Submissions to the WMT19 News Translation Task. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Brian Thompson, Jeremy Gwinnup, Huda Khayrallah, Kevin Duh, and Philipp Koehn. 2019. Overcoming catastrophic forgetting during domain adaptation of neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2062–2068, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Antonio Toral, Lukas Edman, Galiya Yeshmagambetova, and Jennifer Spenader. 2019. Neural Machine Translation for English–Kazakh with Morphological

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

Andre Tättar, Elizaveta Korotkova, and Mark Fishel. 2019. University of Tartu's Multilingual Multi-domain WMT19 News Translation Shared Task Submission. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

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.

Yiren Wang, Yingce Xia, Tianyu He, Fei Tian, Tao Qin, Cheng Xiang Zhai, and Tie-Yan Liu. 2019. Multi-agent dual learning. In *International Conference on Learning Representations*.

Yingce Xia, Xu Tan, Fei Tian, Fei Gao, Weicong Chen, Yang Fan, Linyuan Gong, Yichong Leng, Renqian Luo, Yiren Wang, Lijun Wu, Jinhua Zhu, Tao QIN, and Tie-Yan Liu. 2019. Microsoft Research Asia's Systems for WMT19. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Jingyi Zhang and Josef van Genabith. 2019. DFKI-NMT Submission to the WMT19 News Translation Task. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

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

Jinhua Zhu, Fei Gao, Lijun Wu, Yingce Xia, Tao Qin, Wengang Zhou, Xueqi Cheng, and Tie-Yan Liu. 2019. Soft contextual data augmentation for neural machine translation. *CoRR*, abs/1905.10523.

	BAIDU-SYSTEM	KSAI-SYSTEM	MSRA	MSRA	NEU	BTRANS	ONLINE-B	BTRANS-ENSEMBLE	UEDIN	ONLINE-Y	NICT	ONLINE-A	ONLINE-G	ONLINE-X	APPRENTICE-C
BAIDU-SYSTEM	-	0.03	0.09⋆	0.10⋆	0.10†	0.11†	0.13‡	0.19‡	0.24‡	0.25‡	0.29‡	0.36‡	0.50‡	0.78‡	1.25‡
KSAI-SYSTEM	-0.03	-	0.06	0.07	0.07	0.08⋆	0.10†	0.16‡	0.21‡	0.22‡	0.27‡	0.33‡	0.47‡	0.75‡	1.22‡
MSRA	-0.09	-0.06	-	0.01	0.01	0.02	0.04	0.10‡	0.15‡	0.15‡	0.20‡	0.27‡	0.41‡	0.69‡	1.16‡
MSRA	-0.10	-0.07	-0.01	-	0.00	0.01	0.03	0.09†	0.14‡	0.15‡	0.19‡	0.26‡	0.40‡	0.68‡	1.15‡
NEU	-0.10	-0.07	-0.01	0.00	-	0.01	0.03	0.09†	0.14‡	0.14‡	0.19‡	0.26‡	0.39‡	0.68‡	1.15‡
BTRANS	-0.11	-0.08	-0.02	-0.01	-0.01	-	0.02	0.08†	0.13‡	0.14‡	0.19‡	0.25‡	0.39‡	0.67‡	1.14‡
ONLINE-B	-0.13	-0.10	-0.04	-0.03	-0.03	-0.02	-	0.06⋆	0.11‡	0.11†	0.16‡	0.23‡	0.36‡	0.64‡	1.12‡
BTRANS-ENSEMBLE	-0.19	-0.16	-0.10	-0.09	-0.09	-0.08	-0.06	-	0.05	0.05	0.10†	0.17‡	0.30‡	0.59‡	1.06‡
UEDIN	-0.24	-0.21	-0.15	-0.14	-0.14	-0.13	-0.11	-0.05	-	0.01	0.05⋆	0.12‡	0.26‡	0.54‡	1.01‡
ONLINE-Y	-0.25	-0.22	-0.15	-0.15	-0.14	-0.14	-0.11	-0.05	-0.01	-	0.05⋆	0.11‡	0.25‡	0.53‡	1.01‡
NICT	-0.29	-0.27	-0.20	-0.19	-0.19	-0.19	-0.16	-0.10	-0.05	-0.05	-	0.07⋆	0.20‡	0.48‡	0.96‡
ONLINE-A	-0.36	-0.33	-0.27	-0.26	-0.26	-0.25	-0.23	-0.17	-0.12	-0.11	-0.07	-	0.14†	0.42‡	0.89‡
ONLINE-G	-0.50	-0.47	-0.41	-0.40	-0.39	-0.39	-0.36	-0.30	-0.26	-0.25	-0.20	-0.14	-	0.28‡	0.76‡
ONLINE-X	-0.78	-0.75	-0.69	-0.68	-0.68	-0.67	-0.64	-0.59	-0.54	-0.53	-0.48	-0.42	-0.28	-	0.47‡
APPRENTICE-C	-1.25	-1.22	-1.16	-1.15	-1.15	-1.14	-1.12	-1.06	-1.01	-1.01	-0.96	-0.89	-0.76	-0.47	-
score	0.29	0.27	0.20	0.20	0.19	0.19	0.16	0.10	0.05	0.05	0.00	-0.07	-0.20	-0.48	-0.96
rank	1–7	1–7	1–7	1–7	1–7	1–7	1–7	8–10	8–10	8–10	11	12	13	14	15

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

A Differences in Human Scores

Tables 32–49 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 ⋆ 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.

	HUMAN	KSAI-SYSTEM	BAIDU-SYSTEM	NEU	ONLINE-A	XZL-NMT	UEDIN	BTRANS	NICT	ONLINE-B	ONLINE-Y	ONLINE-G	ONLINE-X
HUMAN	-	0.06†	0.09‡	0.16‡	0.32‡	0.33‡	0.35‡	0.36‡	0.37‡	0.49‡	0.59‡	0.63‡	0.92‡
KSAI-SYSTEM	-0.06	-	0.03	0.10†	0.25‡	0.26‡	0.29‡	0.30‡	0.31‡	0.43‡	0.52‡	0.57‡	0.86‡
BAIDU-SYSTEM	-0.09	-0.03	-	0.07	0.23‡	0.24‡	0.26‡	0.27‡	0.28‡	0.40‡	0.50‡	0.54‡	0.83‡
NEU	-0.16	-0.10	-0.07	-	0.16‡	0.17‡	0.19‡	0.20‡	0.21‡	0.33‡	0.43‡	0.47‡	0.76‡
ONLINE-A	-0.32	-0.25	-0.23	-0.16	-	0.01	0.04	0.04	0.05	0.18‡	0.27‡	0.31‡	0.60‡
XZL-NMT	-0.33	-0.26	-0.24	-0.17	-0.01	-	0.03	0.03	0.04	0.17‡	0.26‡	0.30‡	0.60‡
UEDIN	-0.35	-0.29	-0.26	-0.19	-0.04	-0.03	-	0.01	0.02⋆	0.14‡	0.23‡	0.28‡	0.57‡
BTRANS	-0.36	-0.30	-0.27	-0.20	-0.04	-0.03	-0.01	-	0.01	0.13‡	0.23‡	0.27‡	0.56‡
NICT	-0.37	-0.31	-0.28	-0.21	-0.05	-0.04	-0.02	-0.01	-	0.12‡	0.22‡	0.26‡	0.55‡
ONLINE-B	-0.49	-0.43	-0.40	-0.33	-0.18	-0.17	-0.14	-0.13	-0.12	-	0.09⋆	0.14‡	0.43‡
ONLINE-Y	-0.59	-0.52	-0.50	-0.43	-0.27	-0.26	-0.23	-0.23	-0.22	-0.09	-	0.04	0.34‡
ONLINE-G	-0.63	-0.57	-0.54	-0.47	-0.31	-0.30	-0.28	-0.27	-0.26	-0.14	-0.04	-	0.29‡
ONLINE-X	-0.92	-0.86	-0.83	-0.76	-0.60	-0.60	-0.57	-0.56	-0.55	-0.43	-0.34	-0.29	-
score	0.37	0.31	0.28	0.21	0.05	0.04	0.02	0.01	0.00	-0.13	-0.22	-0.26	-0.55
rank	1	2–4	2–4	2–4	5–9	5–9	5–9	5–9	5–9	10	11–12	11–12	13

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

	HUMAN	CUNI-DocTransformer-T2T	CUNI-Transformer-T2T-2018	CUNI-Transformer-T2T-2019	CUNI-DocTransformer-Marian	UEDIN	ONLINE-Y	TartuNLP-C	ONLINE-G	ONLINE-B	ONLINE-A	ONLINE-X
HUMAN	-	0.24‡	0.24‡	0.25‡	0.42‡	0.44‡	0.80‡	0.84‡	0.94‡	0.98‡	1.24‡	1.29‡
CUNI-DocTransformer-T2T	-0.24	-	0.00	0.01	0.18‡	0.20‡	0.56‡	0.60‡	0.70‡	0.74‡	1.00‡	1.05‡
CUNI-Transformer-T2T-2018	-0.24	0.00	-	0.01	0.18‡	0.20‡	0.56‡	0.60‡	0.70‡	0.74‡	1.00‡	1.05‡
CUNI-Transformer-T2T-2019	-0.25	-0.01	-0.01	-	0.17‡	0.18‡	0.54‡	0.58‡	0.69‡	0.72‡	0.98‡	1.04‡
CUNI-DocTransformer-Marian	-0.42	-0.18	-0.18	-0.17	-	0.02	0.38‡	0.42‡	0.52‡	0.56‡	0.82‡	0.87‡
UEDIN	-0.44	-0.20	-0.20	-0.18	-0.02	-	0.36‡	0.40‡	0.51‡	0.54‡	0.80‡	0.86‡
ONLINE-Y	-0.80	-0.56	-0.56	-0.54	-0.38	-0.36	-	0.04	0.14†	0.18‡	0.44‡	0.49‡
TartuNLP-C	-0.84	-0.60	-0.60	-0.58	-0.42	-0.40	-0.04	-	0.10⋆	0.14‡	0.40‡	0.46‡
ONLINE-G	-0.94	-0.70	-0.70	-0.69	-0.52	-0.51	-0.14	-0.10	-	0.04⋆	0.29‡	0.35‡
ONLINE-B	-0.98	-0.74	-0.74	-0.72	-0.56	-0.54	-0.18	-0.14	-0.04	-	0.26‡	0.31‡
ONLINE-A	-1.24	-1.00	-1.00	-0.98	-0.82	-0.80	-0.44	-0.40	-0.29	-0.26	-	0.06⋆
ONLINE-X	-1.29	-1.05	-1.05	-1.04	-0.87	-0.86	-0.49	-0.46	-0.35	-0.31	-0.06	-
score	0.64	0.40	0.40	0.39	0.22	0.21	-0.16	-0.20	-0.30	-0.34	-0.59	-0.65
rank	1	2–4	2–4	2–4	5–6	5–6	7–8	7–8	9	10	11	12

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

	FACEBOOK-FAIR	RWTH-AACHEN-SYSTEM	MSRA	ONLINE-B	JHU	MLLP-UPV	DFKI-NMT	UCAM	ONLINE-A	NEU	HUMAN	UEDIN	ONLINE-Y	TARTUNLP-C	ONLINE-G	PROMT-NMT-DE-EN	ONLINE-X
FACEBOOK-FAIR	-	0.01†	0.01	0.03	0.06†	0.08‡	0.08‡	0.08†	0.10‡	0.11‡	0.12‡	0.13‡	0.14‡	0.14‡	0.20‡	0.27‡	0.34‡
RWTH-AACHEN-SYSTEM	-0.01	-	0.00	0.02	0.05	0.07†	0.07	0.07	0.09	0.10	0.11⋆	0.12‡	0.13‡	0.13‡	0.19‡	0.26‡	0.33‡
MSRA	-0.01	0.00⋆	-	0.02	0.05⋆	0.07‡	0.07†	0.07†	0.09‡	0.10‡	0.11‡	0.12‡	0.13‡	0.13‡	0.19‡	0.26‡	0.33‡
ONLINE-B	-0.03	-0.02	-0.02	-	0.03	0.05‡	0.05⋆	0.05	0.07†	0.08†	0.09†	0.11‡	0.11‡	0.11‡	0.17‡	0.25‡	0.31‡
JHU	-0.06	-0.05	-0.05	-0.03	-	0.02⋆	0.02	0.02	0.04	0.05	0.06	0.08†	0.08‡	0.08‡	0.14‡	0.21‡	0.28‡
MLLP-UPV	-0.08	-0.07	-0.07	-0.05	-0.02	-	0.00	0.00	0.02	0.03	0.04	0.06	0.06	0.06	0.12†	0.20‡	0.26‡
DFKI-NMT	-0.08	-0.07	-0.07	-0.05	-0.02	0.00	-	0.00	0.02	0.03	0.04	0.06⋆	0.06⋆	0.06†	0.12‡	0.19‡	0.26‡
UCAM	-0.08	-0.07	-0.07	-0.05	-0.02	0.00	0.00	-	0.02	0.03	0.04	0.05⋆	0.06†	0.06†	0.12‡	0.19‡	0.26‡
ONLINE-A	-0.10	-0.09	-0.09	-0.07	-0.04	-0.02	-0.02	-0.02	-	0.01	0.02	0.04	0.04⋆	0.04⋆	0.10‡	0.18‡	0.24‡
NEU	-0.11	-0.10	-0.10	-0.08	-0.05	-0.03	-0.03	-0.03	-0.01	-	0.01	0.03	0.03⋆	0.03⋆	0.09‡	0.17‡	0.23‡
HUMAN	-0.12	-0.11	-0.11	-0.09	-0.06	-0.04	-0.04	-0.04	-0.02	-0.01	-	0.02	0.02⋆	0.02⋆	0.08‡	0.16‡	0.22‡
UEDIN	-0.13	-0.12	-0.12	-0.11	-0.08	-0.06	-0.06	-0.05	-0.04	-0.03	-0.02	-	0.00	0.00	0.06⋆	0.14‡	0.20‡
ONLINE-Y	-0.14	-0.13	-0.13	-0.11	-0.08	-0.06	-0.06	-0.06	-0.04	-0.03	-0.02	0.00	-	0.00	0.06	0.14‡	0.20‡
TARTUNLP-C	-0.14	-0.13	-0.13	-0.11	-0.08	-0.06	-0.06	-0.06	-0.04	-0.03	-0.02	0.00	0.00	-	0.06	0.13‡	0.20‡
ONLINE-G	-0.20	-0.19	-0.19	-0.17	-0.14	-0.12	-0.12	-0.12	-0.10	-0.09	-0.08	-0.06	-0.06	-0.06	-	0.08‡	0.14‡
PROMT-NMT-DE-EN	-0.27	-0.26	-0.26	-0.25	-0.21	-0.20	-0.19	-0.19	-0.18	-0.17	-0.16	-0.14	-0.14	-0.13	-0.08	-	0.06⋆
ONLINE-X	-0.34	-0.33	-0.33	-0.31	-0.28	-0.26	-0.26	-0.26	-0.24	-0.23	-0.22	-0.20	-0.20	-0.20	-0.14	-0.06	-
score	0.15	0.14	0.14	0.12	0.09	0.07	0.07	0.07	0.05	0.04	0.03	0.01	0.01	0.01	-0.05	-0.13	-0.19
rank	1–3	1–3	1–3	4–15	4–15	4–15	4–15	4–15	4–15	4–15	4–15	4–15	4–15	4–15	4–15	16	17

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

	FACEBOOK-FAIR	MICROSOFT-SENT-DOC	MICROSOFT-DOC	HUMAN	MSRA	UCAM	NEU	MLLP-UPV	ETRANSLATION	DFKI-NMT	ONLINE-B	MICROSOFT-SENT	JHU	HELSINKI-NLP	ONLINE-Y	LMU-CTX-TF-SINGLE-EN-DE	PROMT-NMT-EN-DE	ONLINE-A	ONLINE-G	UDS-DFKI	TARTUNLP-C	ONLINE-X	EN-DE-TASK
FACEBOOK-FAIR	-	0.04‡	0.05‡	0.11‡	0.13‡	0.13‡	0.14‡	0.16‡	0.22‡	0.23‡	0.25‡	0.25‡	0.27‡	0.27‡	0.31‡	0.34‡	0.35‡	0.42‡	0.47‡	0.48‡	0.48‡	0.75‡	2.12‡
MICROSOFT-SENT-DOC	-0.04	-	0.02	0.07	0.10‡	0.10†	0.10‡	0.12‡	0.18‡	0.19‡	0.22‡	0.22‡	0.23‡	0.23‡	0.27‡	0.30‡	0.31‡	0.38‡	0.43‡	0.44‡	0.44‡	0.71‡	2.08‡
MICROSOFT-DOC	-0.05	-0.02	-	0.06	0.08⋆	0.08	0.09⋆	0.11‡	0.17‡	0.18‡	0.20‡	0.20‡	0.21‡	0.22‡	0.26‡	0.29‡	0.29‡	0.37‡	0.41‡	0.42‡	0.43‡	0.70‡	2.06‡
HUMAN	-0.11	-0.07	-0.06	-	0.03⋆	0.03	0.03⋆	0.05‡	0.11‡	0.12‡	0.15‡	0.15‡	0.16‡	0.16‡	0.20‡	0.23‡	0.24‡	0.31‡	0.36‡	0.37‡	0.37‡	0.64‡	2.01‡
MSRA	-0.13	-0.10	-0.08	-0.03	-	0.00	0.01	0.03	0.08⋆	0.10‡	0.12†	0.12⋆	0.13‡	0.14†	0.18‡	0.20‡	0.21‡	0.29‡	0.33‡	0.34‡	0.35‡	0.61‡	1.98‡
UCAM	-0.13	-0.10	-0.08	-0.03	0.00	-	0.00	0.02⋆	0.08†	0.09‡	0.12‡	0.12†	0.13‡	0.14‡	0.17‡	0.20‡	0.21‡	0.29‡	0.33‡	0.34‡	0.34‡	0.61‡	1.98‡
NEU	-0.14	-0.10	-0.09	-0.03	-0.01	0.00	-	0.02	0.08⋆	0.09‡	0.11†	0.11⋆	0.13‡	0.13‡	0.17‡	0.20‡	0.21‡	0.28‡	0.33‡	0.34‡	0.34‡	0.61‡	1.98‡
MLLP-UPV	-0.16	-0.12	-0.11	-0.05	-0.03	-0.02	-0.02	-	0.06	0.07‡	0.09⋆	0.09	0.11‡	0.11⋆	0.15‡	0.18‡	0.19‡	0.26‡	0.31‡	0.32‡	0.32‡	0.59‡	1.96‡
ETRANSLATION	-0.22	-0.18	-0.17	-0.11	-0.08	-0.08	-0.08	-0.06	-	0.01⋆	0.04	0.04	0.05⋆	0.05	0.09‡	0.12†	0.13‡	0.20‡	0.25‡	0.26‡	0.26‡	0.53‡	1.90‡
DFKI-NMT	-0.23	-0.19	-0.18	-0.12	-0.10	-0.09	-0.09	-0.07	-0.01	-	0.02	0.02	0.04	0.04	0.08‡	0.11	0.12‡	0.19‡	0.24‡	0.25‡	0.25‡	0.52‡	1.89‡
ONLINE-B	-0.25	-0.22	-0.20	-0.15	-0.12	-0.12	-0.11	-0.09	-0.04	-0.02	-	0.00	0.01	0.02	0.06‡	0.08⋆	0.09‡	0.17‡	0.21‡	0.22‡	0.23‡	0.49‡	1.86‡
MICROSOFT-SENT	-0.25	-0.22	-0.20	-0.15	-0.12	-0.12	-0.11	-0.09	-0.04	-0.02	0.00	-	0.01⋆	0.02	0.06‡	0.08†	0.09‡	0.17‡	0.21‡	0.22‡	0.23‡	0.49‡	1.86‡
JHU	-0.27	-0.23	-0.21	-0.16	-0.13	-0.13	-0.13	-0.11	-0.05	-0.04	-0.01	-0.01	-	0.00	0.04⋆	0.07	0.08⋆	0.15‡	0.20‡	0.21‡	0.21‡	0.48‡	1.85‡
HELSINKI-NLP	-0.27	-0.23	-0.22	-0.16	-0.14	-0.14	-0.13	-0.11	-0.05	-0.04	-0.02	-0.02	0.00	-	0.04†	0.07	0.08†	0.15‡	0.20‡	0.21‡	0.21‡	0.48‡	1.85‡
ONLINE-Y	-0.31	-0.27	-0.26	-0.20	-0.18	-0.17	-0.17	-0.15	-0.09	-0.08	-0.06	-0.06	-0.04	-0.04	-	0.03	0.04	0.11	0.16‡	0.17†	0.17‡	0.44‡	1.81‡
LMU-CTX-TF-SINGLE-EN-DE	-0.34	-0.30	-0.29	-0.23	-0.20	-0.20	-0.20	-0.18	-0.12	-0.11	-0.08	-0.08	-0.07	-0.07	-0.03	-	0.01	0.08†	0.13‡	0.14‡	0.14‡	0.41‡	1.78‡
PROMT-NMT-EN-DE	-0.35	-0.31	-0.29	-0.24	-0.21	-0.21	-0.21	-0.19	-0.13	-0.12	-0.09	-0.09	-0.08	-0.08	-0.04	-0.01	-	0.07	0.12‡	0.13†	0.13‡	0.40‡	1.77‡
ONLINE-A	-0.42	-0.38	-0.37	-0.31	-0.29	-0.29	-0.28	-0.26	-0.20	-0.19	-0.17	-0.17	-0.15	-0.15	-0.11	-0.08	-0.07	-	0.05‡	0.06	0.06†	0.33‡	1.70‡
ONLINE-G	-0.47	-0.43	-0.41	-0.36	-0.33	-0.33	-0.33	-0.31	-0.25	-0.24	-0.21	-0.21	-0.20	-0.20	-0.16	-0.13	-0.12	-0.05	-	0.01	0.01	0.28‡	1.65‡
UDS-DFKI	-0.48	-0.44	-0.42	-0.37	-0.34	-0.34	-0.34	-0.32	-0.26	-0.25	-0.22	-0.22	-0.21	-0.21	-0.17	-0.14	-0.13	-0.06	-0.01	-	0.00	0.27‡	1.64‡
TARTUNLP-C	-0.48	-0.44	-0.43	-0.37	-0.35	-0.34	-0.34	-0.32	-0.26	-0.25	-0.23	-0.23	-0.21	-0.21	-0.17	-0.14	-0.13	-0.06	-0.01	0.00	-	0.27‡	1.64‡
ONLINE-X	-0.75	-0.71	-0.70	-0.64	-0.61	-0.61	-0.61	-0.59	-0.53	-0.52	-0.49	-0.49	-0.48	-0.48	-0.44	-0.41	-0.40	-0.33	-0.28	-0.27	-0.27	-	1.37‡
EN-DE-TASK	-2.12	-2.08	-2.06	-2.01	-1.98	-1.98	-1.98	-1.96	-1.90	-1.89	-1.86	-1.86	-1.85	-1.85	-1.81	-1.78	-1.77	-1.70	-1.65	-1.64	-1.64	-1.37	-
score	0.35	0.31	0.30	0.24	0.21	0.21	0.21	0.19	0.13	0.12	0.09	0.09	0.08	0.08	0.04	0.01	0.00	-0.07	-0.12	-0.13	-0.13	-0.40	-1.77
rank	1	2–20	2–20	2–20	2–20	2–20	2–20	2–20	2–20	2–20	2–20	2–20	2–20	2–20	2–20	2–20	2–20	2–20	2–20	2–20	21	22	23

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

	MSRA	ONLINE-Y	GTCOM-PRIMARY	USYD	ONLINE-B	HELSINKI-NLP	ONLINE-A	ONLINE-G	TARTUNLP-C	ONLINE-X	PARFDA	APERTIUM-FIN-ENG-UNCONSTRAINED-FIEN
MSRA	-	0.02⋆	0.02⋆	0.04⋆	0.18‡	0.18‡	0.27‡	0.33‡	0.34‡	0.36‡	0.49‡	0.80‡
ONLINE-Y	-0.02	-	0.00	0.02	0.16‡	0.16‡	0.25‡	0.31‡	0.32‡	0.34‡	0.47‡	0.78‡
GTCOM-PRIMARY	-0.02	0.00	-	0.02	0.15‡	0.16‡	0.25‡	0.31‡	0.31‡	0.33‡	0.47‡	0.78‡
USYD	-0.04	-0.02	-0.02	-	0.14‡	0.14‡	0.23‡	0.29‡	0.30‡	0.32‡	0.45‡	0.76‡
ONLINE-B	-0.18	-0.16	-0.15	-0.14	-	0.00	0.09‡	0.15‡	0.16‡	0.18‡	0.32‡	0.62‡
HELSINKI-NLP	-0.18	-0.16	-0.16	-0.14	0.00	-	0.09‡	0.15‡	0.16‡	0.18‡	0.31‡	0.62‡
ONLINE-A	-0.27	-0.25	-0.25	-0.23	-0.09	-0.09	-	0.06	0.06⋆	0.08†	0.22‡	0.53‡
ONLINE-G	-0.33	-0.31	-0.31	-0.29	-0.15	-0.15	-0.06	-	0.01	0.03	0.17‡	0.47‡
TARTUNLP-C	-0.34	-0.32	-0.31	-0.30	-0.16	-0.16	-0.06	-0.01	-	0.02	0.16‡	0.46‡
ONLINE-X	-0.36	-0.34	-0.33	-0.32	-0.18	-0.18	-0.08	-0.03	-0.02	-	0.14‡	0.45‡
PARFDA	-0.49	-0.47	-0.47	-0.45	-0.32	-0.31	-0.22	-0.17	-0.16	-0.14	-	0.31‡
APERTIUM-FIN-ENG-UNCONSTRAINED-FIEN	-0.80	-0.78	-0.78	-0.76	-0.62	-0.62	-0.53	-0.47	-0.46	-0.45	-0.31	-
score	0.28	0.27	0.26	0.24	0.11	0.10	0.01	-0.04	-0.05	-0.07	-0.21	-0.52
rank	1	2–4	2–4	2–4	5–6	5–6	7–10	7–10	7–10	7–10	11	12

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

	HUMAN	GTCOM-PRIMARY	MSRA	ONLINE-Y	NICT	HELSINKI-NLP	ONLINE-G	ONLINE-B	TARTUNLP-C	ONLINE-A	ONLINE-X	HELSINKI-NLP–RULE-BASED-	APERTIUM-FIN-ENG-UNCONSTRAINED-EN-FI
HUMAN	-	0.42‡	0.44‡	0.73‡	0.81‡	0.92‡	0.93‡	0.97‡	1.13‡	1.39‡	1.40‡	1.53‡	2.27‡
GTCOM-PRIMARY	-0.42	-	0.02	0.31‡	0.39‡	0.50‡	0.51‡	0.55‡	0.71‡	0.97‡	0.98‡	1.11‡	1.85‡
MSRA	-0.44	-0.02	-	0.29‡	0.37‡	0.48‡	0.50‡	0.53‡	0.69‡	0.95‡	0.97‡	1.09‡	1.83‡
ONLINE-Y	-0.73	-0.31	-0.29	-	0.08⋆	0.19‡	0.20‡	0.24‡	0.40‡	0.66‡	0.67‡	0.80‡	1.54‡
NICT	-0.81	-0.39	-0.37	-0.08	-	0.11‡	0.13‡	0.16‡	0.32‡	0.58‡	0.60‡	0.72‡	1.46‡
HELSINKI-NLP	-0.92	-0.50	-0.48	-0.19	-0.11	-	0.02	0.05⋆	0.21‡	0.47‡	0.49‡	0.61‡	1.35‡
ONLINE-G	-0.93	-0.51	-0.50	-0.20	-0.13	-0.02	-	0.04	0.20‡	0.46‡	0.47‡	0.59‡	1.33‡
ONLINE-B	-0.97	-0.55	-0.53	-0.24	-0.16	-0.05	-0.04	-	0.16‡	0.42‡	0.43‡	0.56‡	1.30‡
TARTUNLP-C	-1.13	-0.71	-0.69	-0.40	-0.32	-0.21	-0.20	-0.16	-	0.26‡	0.27‡	0.40‡	1.14‡
ONLINE-A	-1.39	-0.97	-0.95	-0.66	-0.58	-0.47	-0.46	-0.42	-0.26	-	0.01	0.14‡	0.88‡
ONLINE-X	-1.40	-0.98	-0.97	-0.67	-0.60	-0.49	-0.47	-0.43	-0.27	-0.01	-	0.12‡	0.86‡
HELSINKI-NLP–RULE-BASED-	-1.53	-1.11	-1.09	-0.80	-0.72	-0.61	-0.59	-0.56	-0.40	-0.14	-0.12	-	0.74‡
APERTIUM-FIN-ENG-UNCONSTRAINED-EN-FI	-2.27	-1.85	-1.83	-1.54	-1.46	-1.35	-1.33	-1.30	-1.14	-0.88	-0.86	-0.74	-
score	1.01	0.59	0.57	0.28	0.20	0.09	0.07	0.04	-0.13	-0.38	-0.40	-0.52	-1.26
rank	1	2–3	2–3	4	5	6–8	6–8	6–8	9	10–11	10–11	12	13

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

	NEU	UEDIN	GTCOM-PRIMARY	CUNI-T2T-TRANSFER-GUEN	AYLIEN-MT-GU-EN-MULTILINGUAL	NICT	ONLINE-G	IITP-MT	UDS-DFKI	IIITH-MT	JU-SAARLAND
NEU	-	0.08†	0.11‡	0.12‡	0.14‡	0.17‡	0.40‡	0.40‡	0.49‡	0.51‡	0.81‡
UEDIN	-0.08	-	0.03	0.04	0.06	0.08⋆	0.31‡	0.32‡	0.40‡	0.42‡	0.72‡
GTCOM-PRIMARY	-0.11	-0.03	-	0.01	0.03	0.06	0.29‡	0.29‡	0.38‡	0.40‡	0.70‡
CUNI-T2T-TRANSFER-GUEN	-0.12	-0.04	-0.01	-	0.02	0.05	0.28‡	0.28‡	0.37‡	0.39‡	0.69‡
AYLIEN-MT-GU-EN-MULTILINGUAL	-0.14	-0.06	-0.03	-0.02	-	0.02	0.25‡	0.26‡	0.34‡	0.36‡	0.66‡
NICT	-0.17	-0.08	-0.06	-0.05	-0.02	-	0.23‡	0.24‡	0.32‡	0.34‡	0.64‡
ONLINE-G	-0.40	-0.31	-0.29	-0.28	-0.25	-0.23	-	0.00	0.09†	0.11†	0.41‡
IITP-MT	-0.40	-0.32	-0.29	-0.28	-0.26	-0.24	0.00	-	0.08†	0.10†	0.41‡
UDS-DFKI	-0.49	-0.40	-0.38	-0.37	-0.34	-0.32	-0.09	-0.08	-	0.02	0.32‡
IIITH-MT	-0.51	-0.42	-0.40	-0.39	-0.36	-0.34	-0.11	-0.10	-0.02	-	0.30‡
JU-SAARLAND	-0.81	-0.72	-0.70	-0.69	-0.66	-0.64	-0.41	-0.41	-0.32	-0.30	-
score	0.21	0.13	0.10	0.09	0.07	0.04	-0.19	-0.19	-0.28	-0.30	-0.60
rank	1	2–6	2–6	2–6	2–6	2–6	7–8	7–8	9–10	9–10	11

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

	HUMAN	ONLINE-B	GTCOM-PRIMARY	MSRA	UEDIN	CUNI-T2T-TRANSFER-ENGU	JU-SAARLAND-CLEAN-NUM-135-BPE	IITP-MT	NICT	ONLINE-G	ONLINE-X	UDS-DFKI
HUMAN	-	0.04⋆	0.10‡	0.38‡	0.40‡	0.45‡	0.78‡	1.16‡	1.17‡	1.19‡	1.20‡	1.42‡
ONLINE-B	-0.04	-	0.07⋆	0.34‡	0.36‡	0.41‡	0.74‡	1.12‡	1.13‡	1.15‡	1.16‡	1.38‡
GTCOM-PRIMARY	-0.10	-0.07	-	0.28‡	0.29‡	0.34‡	0.68‡	1.06‡	1.06‡	1.09‡	1.10‡	1.31‡
MSRA	-0.38	-0.34	-0.28	-	0.01	0.06	0.40‡	0.78‡	0.78‡	0.81‡	0.82‡	1.04‡
UEDIN	-0.40	-0.36	-0.29	-0.01	-	0.05	0.38‡	0.76‡	0.77‡	0.79‡	0.81‡	1.02‡
CUNI-T2T-TRANSFER-ENGU	-0.45	-0.41	-0.34	-0.06	-0.05	-	0.33‡	0.71‡	0.72‡	0.74‡	0.76‡	0.97‡
JU-SAARLAND-CLEAN-NUM-135-BPE	-0.78	-0.74	-0.68	-0.40	-0.38	-0.33	-	0.38‡	0.39‡	0.41‡	0.42‡	0.64‡
IITP-MT	-1.16	-1.12	-1.06	-0.78	-0.76	-0.71	-0.38	-	0.01	0.03	0.04†	0.26‡
NICT	-1.17	-1.13	-1.06	-0.78	-0.77	-0.72	-0.39	-0.01	-	0.02	0.04†	0.25‡
ONLINE-G	-1.19	-1.15	-1.09	-0.81	-0.79	-0.74	-0.41	-0.03	-0.02	-	0.01⋆	0.23‡
ONLINE-X	-1.20	-1.16	-1.10	-0.82	-0.81	-0.76	-0.42	-0.04	-0.04	-0.01	-	0.22‡
UDS-DFKI	-1.42	-1.38	-1.31	-1.04	-1.02	-0.97	-0.64	-0.26	-0.25	-0.23	-0.22	-
score	0.70	0.66	0.60	0.32	0.30	0.25	-0.08	-0.46	-0.47	-0.49	-0.50	-0.72
rank	1	2	3	4–6	4–6	4–6	7	8–10	8–10	8–10	11	12

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

	GTCOM-PRIMARY	TILDE-NC-NMT	NEU	MSRA	TILDE-C-NMT	ONLINE-B	ONLINE-A	TARTUNLP-C	ONLINE-G	JUMT	ONLINE-X
GTCOM-PRIMARY	-	0.02	0.02	0.03	0.03⋆	0.13‡	0.29‡	0.29‡	0.52‡	0.57‡	0.63‡
TILDE-NC-NMT	-0.02	-	0.00	0.01	0.01	0.11‡	0.27‡	0.28‡	0.50‡	0.55‡	0.61‡
NEU	-0.02	0.00	-	0.01	0.01	0.11‡	0.27‡	0.27‡	0.50‡	0.55‡	0.61‡
MSRA	-0.03	-0.01	-0.01	-	0.00	0.10†	0.26‡	0.27‡	0.49‡	0.54‡	0.60‡
TILDE-C-NMT	-0.03	-0.01	-0.01	0.00	-	0.09†	0.26‡	0.26‡	0.49‡	0.54‡	0.60‡
ONLINE-B	-0.13	-0.11	-0.11	-0.10	-0.09	-	0.16‡	0.17‡	0.39‡	0.44‡	0.50‡
ONLINE-A	-0.29	-0.27	-0.27	-0.26	-0.26	-0.16	-	0.00	0.23‡	0.28‡	0.34‡
TARTUNLP-C	-0.29	-0.28	-0.27	-0.27	-0.26	-0.17	0.00	-	0.22‡	0.28‡	0.34‡
ONLINE-G	-0.52	-0.50	-0.50	-0.49	-0.49	-0.39	-0.23	-0.22	-	0.05	0.11†
JUMT	-0.57	-0.55	-0.55	-0.54	-0.54	-0.44	-0.28	-0.28	-0.05	-	0.06†
ONLINE-X	-0.63	-0.61	-0.61	-0.60	-0.60	-0.50	-0.34	-0.34	-0.11	-0.06	-
score	0.23	0.22	0.21	0.21	0.20	0.11	-0.06	-0.06	-0.28	-0.34	-0.40
rank	1–5	1–5	1–5	1–5	1–5	6	7–8	7–8	9–10	9–10	11

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

	HUMAN	TILDE-NC-NMT	MSRA	TILDE-C-NMT	MSRA	GTCOM-PRIMARY	ETRANSLATION	NEU	ONLINE-B	TARTUNLP-C	ONLINE-A	ONLINE-X	ONLINE-G
HUMAN	-	0.63‡	0.63‡	0.75‡	0.76‡	0.86‡	0.98‡	1.07‡	1.08‡	1.40‡	1.64‡	1.68‡	1.82‡
TILDE-NC-NMT	-0.63	-	0.00	0.13⋆	0.13†	0.23‡	0.35‡	0.44‡	0.45‡	0.77‡	1.01‡	1.05‡	1.19‡
MSRA	-0.63	0.00	-	0.13†	0.13†	0.23‡	0.35‡	0.44‡	0.45‡	0.77‡	1.01‡	1.05‡	1.19‡
TILDE-C-NMT	-0.75	-0.13	-0.13	-	0.00	0.11†	0.23‡	0.32‡	0.32‡	0.65‡	0.88‡	0.93‡	1.07‡
MSRA	-0.76	-0.13	-0.13	0.00	-	0.10†	0.22‡	0.31‡	0.32‡	0.64‡	0.88‡	0.92‡	1.06‡
GTCOM-PRIMARY	-0.86	-0.23	-0.23	-0.11	-0.10	-	0.12†	0.21‡	0.22‡	0.54‡	0.77‡	0.82‡	0.96‡
ETRANSLATION	-0.98	-0.35	-0.35	-0.23	-0.22	-0.12	-	0.09†	0.10†	0.42‡	0.66‡	0.70‡	0.84‡
NEU	-1.07	-0.44	-0.44	-0.32	-0.31	-0.21	-0.09	-	0.01	0.33‡	0.57‡	0.61‡	0.75‡
ONLINE-B	-1.08	-0.45	-0.45	-0.32	-0.32	-0.22	-0.10	-0.01	-	0.32‡	0.56‡	0.60‡	0.74‡
TARTUNLP-C	-1.40	-0.77	-0.77	-0.65	-0.64	-0.54	-0.42	-0.33	-0.32	-	0.24‡	0.28‡	0.42‡
ONLINE-A	-1.64	-1.01	-1.01	-0.88	-0.88	-0.77	-0.66	-0.57	-0.56	-0.24	-	0.05	0.19‡
ONLINE-X	-1.68	-1.05	-1.05	-0.93	-0.92	-0.82	-0.70	-0.61	-0.60	-0.28	-0.05	-	0.14†
ONLINE-G	-1.82	-1.19	-1.19	-1.07	-1.06	-0.96	-0.84	-0.75	-0.74	-0.42	-0.19	-0.14	-
score	1.02	0.39	0.39	0.26	0.26	0.15	0.04	-0.05	-0.06	-0.38	-0.62	-0.67	-0.81
rank	1	2–3	2–3	4–5	4–5	6	7	8–9	8–9	10	11–12	11–12	13

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

	ONLINE-B	NEU	RUG-KKEN-MORFESSOR	ONLINE-G	TALP-UPC-2019-KKEN	NRC-CNRC	FRANK-S-MT	NICT	CUNI-T2T-TRANSFER-KKEN	UMD	DBMS-KU-KKEN
ONLINE-B	-	0.05	0.08⋆	0.14‡	0.16‡	0.18‡	0.20‡	0.21‡	0.27‡	0.75‡	1.33‡
NEU	-0.05	-	0.03	0.08†	0.10‡	0.13‡	0.15‡	0.15‡	0.21‡	0.69‡	1.28‡
RUG-KKEN-MORFESSOR	-0.08	-0.03	-	0.06⋆	0.08⋆	0.10‡	0.12†	0.12‡	0.19‡	0.67‡	1.25‡
ONLINE-G	-0.14	-0.08	-0.06	-	0.02	0.04	0.07	0.07⋆	0.13‡	0.61‡	1.19‡
TALP-UPC-2019-KKEN	-0.16	-0.10	-0.08	-0.02	-	0.02	0.05	0.05	0.11†	0.59‡	1.17‡
NRC-CNRC	-0.18	-0.13	-0.10	-0.04	-0.02	-	0.03	0.03	0.09⋆	0.57‡	1.15‡
FRANK-S-MT	-0.20	-0.15	-0.12	-0.07	-0.05	-0.03	-	0.00	0.06⋆	0.54‡	1.12‡
NICT	-0.21	-0.15	-0.12	-0.07	-0.05	-0.03	0.00	-	0.06	0.54‡	1.12‡
CUNI-T2T-TRANSFER-KKEN	-0.27	-0.21	-0.19	-0.13	-0.11	-0.09	-0.06	-0.06	-	0.48‡	1.06‡
UMD	-0.75	-0.69	-0.67	-0.61	-0.59	-0.57	-0.54	-0.54	-0.48	-	0.58‡
DBMS-KU-KKEN	-1.33	-1.28	-1.25	-1.19	-1.17	-1.15	-1.12	-1.12	-1.06	-0.58	-
score	0.27	0.22	0.19	0.13	0.11	0.09	0.07	0.06	0.00	-0.48	-1.06
rank	1–3	1–3	1–3	4–9	4–9	4–9	4–9	4–9	4–9	10	11

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

	HUMAN	UALACANT—NMT	ONLINE-B	UALACANT—N	RBMT	NEU	MSRA	CUNI-T2T-TRANSFER-ENKK	ONLINE-G	RUG-ENKK-BPE	TALP-UPC-2019-ENKK	NICT	DBMS-KU-ENKK
HUMAN	-	0.48‡	0.50‡	0.52‡	0.52‡	0.62‡	0.64‡	0.65‡	0.67‡	0.79‡	1.07‡	1.24‡	2.14‡
UALACANT—NMT	-0.48	-	0.02	0.04	0.04	0.14‡	0.15‡	0.16‡	0.18‡	0.31‡	0.59‡	0.75‡	1.66‡
ONLINE-B	-0.50	-0.02	-	0.02	0.02	0.12†	0.14†	0.15†	0.17‡	0.29‡	0.57‡	0.74‡	1.64‡
UALACANT—N	-0.52	-0.04	-0.02	-	0.00	0.10†	0.11†	0.13⋆	0.14‡	0.27‡	0.55‡	0.72‡	1.62‡
RBMT	-0.52	-0.04	-0.02	0.00	-	0.10†	0.11†	0.13⋆	0.14‡	0.27‡	0.55‡	0.72‡	1.62‡
NEU	-0.62	-0.14	-0.12	-0.10	-0.10	-	0.02	0.03	0.05	0.18‡	0.45‡	0.62‡	1.52‡
MSRA	-0.64	-0.15	-0.14	-0.11	-0.11	-0.02	-	0.01	0.03	0.16‡	0.44‡	0.60‡	1.50‡
CUNI-T2T-TRANSFER-ENKK	-0.65	-0.16	-0.15	-0.13	-0.13	-0.03	-0.01	-	0.02	0.15‡	0.42‡	0.59‡	1.49‡
ONLINE-G	-0.67	-0.18	-0.17	-0.14	-0.14	-0.05	-0.03	-0.02	-	0.13†	0.41‡	0.57‡	1.47‡
RUG-ENKK-BPE	-0.79	-0.31	-0.29	-0.27	-0.27	-0.18	-0.16	-0.15	-0.13	-	0.28‡	0.44‡	1.35‡
TALP-UPC-2019-ENKK	-1.07	-0.59	-0.57	-0.55	-0.55	-0.45	-0.44	-0.42	-0.41	-0.28	-	0.17‡	1.07‡
NICT	-1.24	-0.75	-0.74	-0.72	-0.72	-0.62	-0.60	-0.59	-0.57	-0.44	-0.17	-	0.90‡
DBMS-KU-ENKK	-2.14	-1.66	-1.64	-1.62	-1.62	-1.52	-1.50	-1.49	-1.47	-1.35	-1.07	-0.90	-
score	0.75	0.26	0.24	0.22	0.22	0.13	0.11	0.10	0.08	-0.05	-0.33	-0.49	-1.40
rank	1	2–5	2–5	2–5	2–5	6–9	6–9	6–9	6–9	10	11	12	13

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

	FACEBOOK-FAIR	ONLINE-G	ETRANSLATION	ONLINE-B	NEU	MSRA	RERANK-RE	ONLINE-Y	ONLINE-A	AFRL-SYSCOMB19	AFRL-EWC	TARTUNLP-U	ONLINE-X	NICT
FACEBOOK-FAIR	-	0.02⋆	0.03⋆	0.04	0.04	0.05⋆	0.07‡	0.08‡	0.13‡	0.14‡	0.20‡	0.20‡	0.25‡	0.46‡
ONLINE-G	-0.02	-	0.01	0.01	0.02	0.03	0.05⋆	0.06	0.11†	0.12‡	0.17‡	0.17‡	0.23‡	0.44‡
ETRANSLATION	-0.03	-0.01	-	0.00	0.01	0.02	0.04	0.05	0.09⋆	0.11‡	0.16‡	0.16‡	0.22‡	0.43‡
ONLINE-B	-0.04	-0.01	0.00	-	0.01	0.02	0.04⋆	0.04⋆	0.09†	0.11‡	0.16‡	0.16‡	0.22‡	0.42‡
NEU	-0.04	-0.02	-0.01	-0.01	-	0.01	0.03⋆	0.04⋆	0.09†	0.10‡	0.15‡	0.15‡	0.21‡	0.42‡
MSRA	-0.05	-0.03	-0.02	-0.02	-0.01	-	0.02	0.03	0.07⋆	0.09†	0.14‡	0.14‡	0.20‡	0.41‡
RERANK-RE	-0.07	-0.05	-0.04	-0.04	-0.03	-0.02	-	0.01	0.05	0.07⋆	0.12‡	0.12‡	0.18‡	0.39‡
ONLINE-Y	-0.08	-0.06	-0.05	-0.04	-0.04	-0.03	-0.01	-	0.05	0.06⋆	0.12‡	0.12‡	0.17‡	0.38‡
ONLINE-A	-0.13	-0.11	-0.09	-0.09	-0.09	-0.07	-0.05	-0.05	-	0.02	0.07†	0.07†	0.13‡	0.33‡
AFRL-SYSCOMB19	-0.14	-0.12	-0.11	-0.11	-0.10	-0.09	-0.07	-0.06	-0.02	-	0.05⋆	0.05	0.11‡	0.32‡
AFRL-EWC	-0.20	-0.17	-0.16	-0.16	-0.15	-0.14	-0.12	-0.12	-0.07	-0.05	-	0.00	0.06†	0.26‡
TARTUNLP-U	-0.20	-0.17	-0.16	-0.16	-0.15	-0.14	-0.12	-0.12	-0.07	-0.05	0.00	-	0.06†	0.26‡
ONLINE-X	-0.25	-0.23	-0.22	-0.22	-0.21	-0.20	-0.18	-0.17	-0.13	-0.11	-0.06	-0.06	-	0.21‡
NICT	-0.46	-0.44	-0.43	-0.42	-0.42	-0.41	-0.39	-0.38	-0.33	-0.32	-0.26	-0.26	-0.21	-
score	0.16	0.13	0.12	0.12	0.12	0.10	0.08	0.08	0.03	0.01	-0.04	-0.04	-0.10	-0.30
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	13	14

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

	HUMAN	FACEBOOK-FAIR	USTC-MCC	ONLINE-G	ONLINE-B	NEU	PROMT-NMT-EN-RU	ONLINE-Y	RERANK-ER	ONLINE-A	TARTUNLP-U	ONLINE-X	NICT
HUMAN	-	0.03	0.20‡	0.26‡	0.27‡	0.31‡	0.32‡	0.38‡	0.72‡	0.80‡	0.85‡	0.90‡	1.81‡
FACEBOOK-FAIR	-0.03	-	0.17‡	0.23‡	0.24‡	0.28‡	0.29‡	0.35‡	0.69‡	0.77‡	0.82‡	0.87‡	1.78‡
USTC-MCC	-0.20	-0.17	-	0.05†	0.06‡	0.11‡	0.11‡	0.18‡	0.52‡	0.60‡	0.64‡	0.69‡	1.60‡
ONLINE-G	-0.26	-0.23	-0.05	-	0.01	0.06⋆	0.06†	0.12‡	0.47‡	0.55‡	0.59‡	0.64‡	1.55‡
ONLINE-B	-0.27	-0.24	-0.06	-0.01	-	0.05	0.05⋆	0.11‡	0.46‡	0.54‡	0.58‡	0.63‡	1.54‡
NEU	-0.31	-0.28	-0.11	-0.06	-0.05	-	0.00	0.07†	0.41‡	0.49‡	0.53‡	0.59‡	1.49‡
PROMT-NMT-EN-RU	-0.32	-0.29	-0.11	-0.06	-0.05	0.00	-	0.06⋆	0.41‡	0.49‡	0.53‡	0.58‡	1.49‡
ONLINE-Y	-0.38	-0.35	-0.18	-0.12	-0.11	-0.07	-0.06	-	0.34‡	0.42‡	0.47‡	0.52‡	1.43‡
RERANK-ER	-0.72	-0.69	-0.52	-0.47	-0.46	-0.41	-0.41	-0.34	-	0.08‡	0.12‡	0.17‡	1.08‡
ONLINE-A	-0.80	-0.77	-0.60	-0.55	-0.54	-0.49	-0.49	-0.42	-0.08	-	0.04	0.09‡	1.00‡
TARTUNLP-U	-0.85	-0.82	-0.64	-0.59	-0.58	-0.53	-0.53	-0.47	-0.12	-0.04	-	0.05‡	0.96‡
ONLINE-X	-0.90	-0.87	-0.69	-0.64	-0.63	-0.59	-0.58	-0.52	-0.17	-0.09	-0.05	-	0.91‡
NICT	-1.81	-1.78	-1.60	-1.55	-1.54	-1.49	-1.49	-1.43	-1.08	-1.00	-0.96	-0.91	-
score	0.54	0.51	0.33	0.28	0.27	0.22	0.22	0.16	-0.19	-0.27	-0.31	-0.36	-1.27
rank	1–2	1–2	3	4–7	4–7	4–7	4–7	8	9	10–11	10–11	12	13

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

	ONLINE-Y	ONLINE-B	NICT	ONLINE-G	NEU-KINGSOFT	ONLINE-A	LMU-UNSUP-NMT-DE-CS	CUNI-UNSUPERVISED-NER-POST	UNSUPERVISED	UNSUPERVISED	CAIRE
ONLINE-Y	-	0.04	0.06⋆	0.11‡	0.25‡	0.29‡	0.53‡	0.59‡	0.75‡	0.83‡	1.23‡
ONLINE-B	-0.04	-	0.02	0.07⋆	0.21‡	0.25‡	0.49‡	0.55‡	0.71‡	0.79‡	1.19‡
NICT	-0.06	-0.02	-	0.05	0.19‡	0.23‡	0.47‡	0.53‡	0.69‡	0.77‡	1.17‡
ONLINE-G	-0.11	-0.07	-0.05	-	0.14‡	0.19‡	0.42‡	0.48‡	0.65‡	0.72‡	1.13‡
NEU-KINGSOFT	-0.25	-0.21	-0.19	-0.14	-	0.05	0.28‡	0.34‡	0.51‡	0.58‡	0.99‡
ONLINE-A	-0.29	-0.25	-0.23	-0.19	-0.05	-	0.23‡	0.30‡	0.46‡	0.54‡	0.94‡
LMU-UNSUP-NMT-DE-CS	-0.53	-0.49	-0.47	-0.42	-0.28	-0.23	-	0.07⋆	0.23‡	0.31‡	0.71‡
CUNI-UNSUPERVISED-NER-POST	-0.59	-0.55	-0.53	-0.48	-0.34	-0.30	-0.07	-	0.16‡	0.24‡	0.64‡
UNSUPERVISED	-0.75	-0.71	-0.69	-0.65	-0.51	-0.46	-0.23	-0.16	-	0.08⋆	0.48‡
UNSUPERVISED	-0.83	-0.79	-0.77	-0.72	-0.58	-0.54	-0.31	-0.24	-0.08	-	0.40‡
CAIRE	-1.23	-1.19	-1.17	-1.13	-0.99	-0.94	-0.71	-0.64	-0.48	-0.40	-
score	0.43	0.39	0.37	0.32	0.18	0.13	-0.10	-0.17	-0.33	-0.41	-0.81
rank	1–4	1–4	1–4	1–4	5–6	5–6	7	8	9	10	11

Table 47: Head to head comparison for German→Czech systems

	MSRA	MLLP-UPV	KYOTO-UNIVERSITY-T2T	LINGUA-CUSTODIA-PRIMARY	LIUM	ONLINE-B	ONLINE-Y	TARTUNLP-C	ONLINE-A	ONLINE-G	ONLINE-X
MSRA	-	0.02	0.09	0.09⋆	0.12⋆	0.21‡	0.23‡	0.35‡	0.44‡	0.49‡	0.71‡
MLLP-UPV	-0.02	-	0.07	0.07	0.10⋆	0.19‡	0.21‡	0.33‡	0.42‡	0.47‡	0.69‡
KYOTO-UNIVERSITY-T2T	-0.09	-0.07	-	0.00	0.04	0.13†	0.15†	0.27‡	0.36‡	0.40‡	0.62‡
LINGUA-CUSTODIA-PRIMARY	-0.09	-0.07	0.00	-	0.03	0.12†	0.14†	0.26‡	0.35‡	0.40‡	0.62‡
LIUM	-0.12	-0.10	-0.04	-0.03	-	0.09†	0.11†	0.23‡	0.32‡	0.37‡	0.58‡
ONLINE-B	-0.21	-0.19	-0.13	-0.12	-0.09	-	0.02	0.14⋆	0.23‡	0.28‡	0.49‡
ONLINE-Y	-0.23	-0.21	-0.15	-0.14	-0.11	-0.02	-	0.12⋆	0.21‡	0.26‡	0.47‡
TARTUNLP-C	-0.35	-0.33	-0.27	-0.26	-0.23	-0.14	-0.12	-	0.09	0.14	0.35‡
ONLINE-A	-0.44	-0.42	-0.36	-0.35	-0.32	-0.23	-0.21	-0.09	-	0.05	0.26‡
ONLINE-G	-0.49	-0.47	-0.40	-0.40	-0.37	-0.28	-0.26	-0.14	-0.05	-	0.22‡
ONLINE-X	-0.71	-0.69	-0.62	-0.62	-0.58	-0.49	-0.47	-0.35	-0.26	-0.22	-
score	0.25	0.23	0.16	0.16	0.13	0.04	0.02	-0.10	-0.19	-0.24	-0.46
rank	1–5	1–5	1–5	1–5	1–5	6–7	6–7	8–10	8–10	8–10	11

Table 48: Head to head comparison for German→French systems

	MSRA	eTRANSLATION	LIUM	MLLP-UPV	ONLINE-Y	ONLINE-G	ONLINE-B	ONLINE-A	TARTUNLP-C	ONLINE-X
MSRA	-	0.02	0.19†	0.23‡	0.27‡	0.29‡	0.30‡	0.31‡	0.37‡	0.68‡
eTRANSLATION	-0.02	-	0.16⋆	0.21⋆	0.24‡	0.26‡	0.28‡	0.29‡	0.34‡	0.66‡
LIUM	-0.19	-0.16	-	0.04	0.08⋆	0.10	0.12⋆	0.12⋆	0.18‡	0.49‡
MLLP-UPV	-0.23	-0.21	-0.04	-	0.04	0.06	0.07	0.08⋆	0.14†	0.45‡
ONLINE-Y	-0.27	-0.24	-0.08	-0.04	-	0.02	0.03	0.04	0.10	0.41‡
ONLINE-G	-0.29	-0.26	-0.10	-0.06	-0.02	-	0.02	0.02	0.08⋆	0.39‡
ONLINE-B	-0.30	-0.28	-0.12	-0.07	-0.03	-0.02	-	0.01	0.06	0.38‡
ONLINE-A	-0.31	-0.29	-0.12	-0.08	-0.04	-0.02	-0.01	-	0.06	0.37‡
TARTUNLP-C	-0.37	-0.34	-0.18	-0.14	-0.10	-0.08	-0.06	-0.06	-	0.31‡
ONLINE-X	-0.68	-0.66	-0.49	-0.45	-0.41	-0.39	-0.38	-0.37	-0.31	-
score	0.27	0.25	0.08	0.04	0.00	-0.02	-0.03	-0.04	-0.10	-0.41
rank	1–2	1–2	3–9	3–9	3–9	3–9	3–9	3–9	3–9	10

Table 49: Head to head comparison for French→German systems