View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Edinburgh Research Explorer

THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Findings of the 2018 Conference on Machine Translation (WMT18)

Citation for published version:
Bojar, O, Federmann, C, Fishel, M, Graham, Y, Haddow, B, Huck, M, Koehn, P & Monz, C 2018, Findings of the 2018 Conference on Machine Translation (WMT18). in Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers. Association for Computational Linguistics, Belgium, Brussels, pp. 272-307, EMNLP 2018 Third Conference on Machine Translation (WMT18), Brussels, Belgium, 31/10/18. DOI: 10.18653/v1/W18-64028

Digital Object Identifier (DOI):
10.18653/v1/W18-64028

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 05. Apr. 2019

Findings of the 2018 Conference on Machine Translation (WMT18)

Ondřej Bojar
Charles University

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

Christof Monz
University of Amsterdam

Abstract

This paper presents the results of the premier shared task organized alongside the Conference on Machine Translation (WMT) 2018. Participants were asked to build machine translation systems for any of 7 language pairs in both directions, to be evaluated on a test set of news stories. The main metric for this task is human judgment of translation quality. This year, we also opened up the task to additional test suites to probe specific aspects of translation.

1 Introduction

The Third Conference on Machine Translation (WMT) held at EMNLP 2018[1] hosts a number of shared tasks on various aspects of machine translation. This conference builds on twelve 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, 2016a, 2017).

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

- biomedical translation (Neves et al., 2018),
- multimodal machine translation (Barrault et al., 2018),
- metrics (Ma et al., 2018),
- quality estimation (Specia et al., 2018),
- automatic post-editing (Chatterjee et al., 2018), and
- parallel corpus filtering (Koehn et al., 2018b).

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 14 translation tasks this year, between English and each of Chinese, Czech, Estonian, German, Finnish, Russian, and Turkish. The Estonian-English language pair was new this year. Similarly to Latvian, which we had covered in 2017, Estonian is a lesser resourced data condition on a challenging language pair. System outputs for each task were evaluated both automatically and manually.

This year the news translation task had two additional sub-tracks: multilingual and unsupervised MT. Both sub-tracks were included into the general list of news translation submissions and are described in more detail in 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-driven machine translation, automatic evaluation, or prediction of translation quality. News transla-

[1] http://www.statmt.org/wmt18/

[2] http://statmt.org/wmt18/results.html

Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 272–307
Brussels, Belgium, October 31 - Novermber 1, 2018. ©2018 Association for Computational Linguistics
https://doi.org/10.18653/v1/W18-64028

tions 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, Russian, and Turkish. A new language this year is Estonian.

We created a test set for each language pair by translating newspaper articles and provided training data.

2.1 Test Data

The test data for this year's task was selected from online sources, as in previous years. We took about 1500 English sentences and translated them into the other languages, and then additional 1500 sentences from each of the other languages and translated them into English. This gave us test sets of about 3000 sentences for our English-X language pairs, which have been either originally written in English and translated into X, or vice versa. 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 stories were translated by professional translators[3] funded by the EU Horizon 2020 projects CRACKER and QT21 (German, Czech), by Yandex,[4] a Russian search engine company (Turkish, Russian), by BAULT, a research community on building and using language technology funded by the University of Helsinki (Finnish) and the University of Tartu (Estonian). The Chinese–English task was sponsored by Nanjing University, Xiamen University, the Institutes of Computing Technology and of Automation, Chinese Academy of Science, Northeastern University (China) and Datum Data Co., Ltd. All of the translations were done directly, and not via an intermediate language.

Since Estonian-English was run for the first time, both the test and development set had to be translated: the size of both was 2000 sentences (4000 in total).

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. Some training corpora were identical from last year (Europarl,[5] Common Crawl, SETIMES2, Russian-English parallel data provided by Yandex, Wikipedia Headlines provided by CMU) and some were updated (United Nations, CzEng,[6] News Commentary v13, monolingual news data). The new Estonian-English language pair had parallel data from Europarl, EU Press Releases and ParaCrawl, as well as a monolingual corpus of mostly news articles called BigEst.[7]

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

2.3 Multilingual and Unsupervised Sub-tracks

This year the news translation task included two sub-tracks: one on multilingual translation and another one on unsupervised MT.

The multilingual sub-track covered any submissions that used any data (monolingual or parallel) from a third language to help translating the language pair in question: for example, using English-Finnish data to improve English-Estonian translation. All entries to this sub-track had to use only the WMT-provided data sets, and thus had to be constrained.

In the unsupervised MT sub-track the participants were further constrained to using only the

[3] In particular, the Czech and German test sets were translated to/from English by the professional level of service of Translated.net, preserving 1-1 segment translation and aiming for literal translation where possible. Each language combination included 2 different translators: the first translator took care of the translation, the second translator was asked to evaluate a representative part of the work to give a score to the first translator. All translators translate towards their mother tongue only and need to provide a proof or their education or professional experience, or to take a test; they are continuously evaluated to understand how they perform on the long term. The domain knowledge of the translators is ensured by matching translators and the documents using T-Rank, http://www.translated.net/en/T-Rank.

[4] http://www.yandex.com/

[5] As of Fall 2011, the proceedings of the European Parliament are no longer translated into all official languages.

[6] WMT18 recommended to use CzEng v1.7 which is a filtered subset of the previous v1.6 (Bojar et al., 2016b), see http://ufal.mff.cuni.cz/czeng/czeng17.

[7] http://statmt.ut.ee/

Europarl Parallel Corpus

	German ↔ English		Czech ↔ English		Finnish ↔ English		Estonian ↔ English	
Sentences	1,920,209		646,605		1,926,114		652,944	
Words	50,486,398	53,008,851	14,946,399	17,376,433	37,814,266	52,723,296	13,033,918	17,453,613
Distinct words	381,583	115,966	172,461	63,039	693,963	115,896	298,021	63,432

News Commentary Parallel Corpus

	German ↔ English		Czech ↔ English		Russian ↔ English		Chinese ↔ English	
Sentences	284,246		218,384		235,159		252,777	
Words	7,243,776	7,174,644	4,942,255	5,411,117	6,230,738	6,230,738	–	6,428,459
Distinct words	182,059	75,590	166,173	66,054	71,021	71,021	–	70,092

Common Crawl Parallel Corpus

	German ↔ English		Czech ↔ English		Russian ↔ English	
Sentences	2,399,123		161,838		878,386	
Words	54,575,405	58,870,638	3,529,783	3,927,378	21,018,793	21,535,122
Distinct words	1,640,835	823,480	210,170	128,212	764,203	432,062

ParaCrawl Parallel Corpus

	German ↔ English		Czech ↔ English		Estonian ↔ English	
Sentences	36,351,593		10,020,250		1,298,103	
Words	595,027,749	623,361,284	116,797,931	122,699,058	37,887,435	39,060,095
Distinct Words	8065519	5,371,211	1,912,633	1,538,696	1,025,961	894,357

	Finnish ↔ English		Russian ↔ English	
Sentences	624,058		1,2061,155	
Words	8,636,936	11,123,014	182,229,052	210,751,004
Distinct Words	379,958	127,006	3,164,200	2,415,633

EU Press Release Parallel Corpus

	German ↔ English		Finnish ↔ English		Estonian ↔ English	
Sentences	1,329,041		583,223		226978	
Words	25,048,312	25,777,997	8,052,607	11,244,602	3,940,058	177,723
Distinct words	398,477	168,725	315,394	94,979	5,209,544	57,059

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

Wiki Headlines Parallel Corpus

	Russian ↔ English		Finnish ↔ English	
Sentences	514,859		153,728	
Words	1,191,474	1,230,644	269,429	354,362
Distinct	282,989	251,328	127,576	96,732

CzEng v1.6 Parallel Corpus

	Czech ↔ English	
Sentences	61,243,252	
Words	737,434,097	3,650,518
Distinct	835,192,627	2,580,902

SE Times 2 Parallel Corpus

	Turkish ↔ English	
Sentences	207,678	
Words	4,626,277	5,147,769
Distinct	155,479	69,927

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.

Language	Sources (Number of Documents)
English	ABC News (1), BBC (4), Brisbane Times (1), CBS News (1), Daily Mail (4), Euronews (3), Globe and Mail (1), Guardian (4), Independent (4), Los Angeles Times (4), MSNBC (3), Novinte (2), New York Times (2), Reuters (3), Russia Today (2), Scotsman (2), Sydney Morning Herald (2), Telegraph (2), The Local (2), Time Magazine (2), UPI (1), Washington Post (3)
Czech	blesk.cz (16), deník.cz (5), Deník Referendum (1), DNES.cz (7), lidovky.cz (6), Novinky.cz (3), Reflex (2), tyden.cz (12), ZDN (2)
German	Aachener Nachrichten (1), Abendzeitung Nürnberg (2), Braunschweiger Zeitung (1), Der Standard (1), Die Presse (1), Euronews (1), Fehmarn24 (1), Handelsblatt (1), Hannoversche Allgemeine (2), Hessische/Niedersächsische Allgemeine (1), In Franken (4), Kreiszeitung (2), Krone (1), Mainpost (1), Merkur (3), Morgenpost (1), n-tv (1), Neue Westfälische (1), oe24 (2), Peiner Allgemeine (1), Passauer Neue Presse (2), Rheinzeitung (1), Rundschau (1), Schwarzwälder Bote (16), Segeberger Zeitung (2), Südkurier (1), Thüringer Allgemeine (1), Thüringer Landeszeitung (1), Volksblatt (2), Volksfreund (3), Westfälische Nachrichten (1), Westdeutsche Zeitung (8).
Estonian	Arileht (7), Maaleht (3), Postimees (17), Sloleht (23).
Finnish	Etelä-Saimaa (2), Etelä-Suomen Sanomat (3), Helsingin Sanomat (4), Iltalehti (13), Ilta-Sanomat (29), Kaleva (12), Kansan Uutiset (1), Karjalainen (13), Kouvolan Sanomat (2).
Russian	aif (4), Altapress (1), Argumenti (19), ERR.ee (3), eg-online.ru (2), Euronews (2), Fakty (5), Infox (2), Izvestiya (25), Kommersant (16), Lenta (9), lgng (3), MK RU (5), nov-pravda.ru (1), pnp.ru (6), rg.ru (4), Vedomosti (3), Versia (1), Vesti (3), zr.ru (1)
Turkish	Hürriyet.com (48), Sabah (96), Sözcü (19)

Table 1: Composition of the test set. 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.

BigEst Estonian Corpus

Sentences	40,404,948
Words	579,221,489
Distinct words	8,134,555

News Language Model Data

	English	German	Czech	Russian	Finnish	Turkish	Estonian
Sentences	192,988,741	260,754,881	66,517,569	39,519,008	14,575,981	4,753,928	817,472
Words	4,428,839,473	4,627,780,738	1,094,215,341	724,582,848	184,523,981	79,067,739	12,880,832
Distinct words	6,468,049	20,276,165	4,269,005	3,397,828	4,391,543	1,025,791	653,980

Common Crawl Language Model Data

	English	German	Czech	Russian	Finnish	Estonian	Turkish	Chinese
Sent.	3,074,921,453	2,872,785,485	333,498,145	1,168,529,851	157,264,161	100,779,314	511,196,951	1,672,324,647
Words	65,128,419,540	65,154,042,103	6,694,811,063	23,313,060,950	2,935,402,545	2,906,100,138	11,882,126,872	–
Dist.	342,760,462	339,983,035	50,162,437	101,436,673	47,083,545	27,618,190	88,463,295	–

Test Set

	Czech ↔ EN		German ↔ EN		Finnish ↔ EN		Estonian ↔ EN	
Sentences.	2983		2998		3000		2000	
Words	47,229	55,920	54,933	58,628	38,149	54,790	30,531	40,158
Distinct words	18,325	12,548	15,996	13,431	17,825	12,043	14,185	10,096

	Russian ↔ EN		Turkish ↔ EN		Chinese ↔ EN	
Sentences.	3000		3000		3981	
Words	51,988	62,925	45,944	60,232	–	98,308
Distinct words	21,116	13,584	19,200	13,444	–	16,955

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.

monolingual training data from WMT; this additionally excluded the monolingual corpora that are largely parallel (monolingual parts of Europarl and News Commentary). The aim of this task was to see how far can one get in terms of translation quality without any parallel data used for training.[8]

While there was no restriction in terms of language pairs, three language pairs were "verbally endorsed": English to/from Turkish, Estonian and German. The motivation behind the choice of languages was to test the effect of multilingual and unsupervised methods on low-resource language pairs (Turkish-English, Estonian-English) and to contrast the results with a resource-rich pair (German-English).

Submissions to both sub-tracks are joined with the main translation track and evaluated without separation in the same way.

2.4 Submitted Systems

We received 103 submissions from 32 institutions. The participating institutions, organized into 35 teams are listed in Table 2 and detailed in the rest of this section. Each system did not necessarily appear in all translation tasks. We also included 39 online MT systems (originating from 5 services), which we anonymized as ONLINE-A,B,F,G,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.

2.4.1 AALTO (Grönroos et al., 2018)

Aalto participated in the constrained condition of the multi-lingual subtrack, with a single system trained to translate from English to both Finnish and Estonian. The system is based on the Transformer (Vaswani et al., 2017) implementation in OpenNMT-py (Klein et al., 2017). It is trained on filtered parallel and filtered back-translated monolingual data. The main contribution is a novel cross-lingual Morfessor (Virpioja et al., 2013) segmentation using cognates extracted from the parallel data. The aim is to improve the consistency of the morphological segmentation. Aalto decode using an ensemble of 3 (et) or 8 (fi) models.

[8] As an exception it was allowed to use a parallel dev set for parameter tuning and/or model selection

2.4.2 AFRL (Gwinnup et al., 2018)

AFRL-SYSCOMB is a system-combination entry consisting of three inputs. The first is an OpenNMT system trained on the provided parallel data except ParaCrawl and the backtranslated corpus used in the AFRL WMT17 system (Gwinnup et al., 2017). This system uses a standard RNN architecture and was fine-tuned with the other available news task test sets. The second is a Marian (Junczys-Dowmunt et al., 2018) system ensembling 5 Univ. Edinburgh "bi-deep" and 6 transformer models all trained on the WMT18 bitexts provided, including ParaCrawl. Some models employed pretrained word embeddings built on BPE'd corpora (Sennrich et al., 2016b). A Marian transformer model performed right-to-left rescoring for this system. The third system is trained with Moses (Koehn et al., 2007), using the same data as the Marian system. Hierarchical reordering and Operation Sequence Model were employed. The 5-gram English language model was trained with KenLM (Heafield, 2011) on the same corpus as the AFRL WMT15 system with the same BPE used in the Marian systems. Lastly, RWTH Jane's system combination (Freitag et al., 2014) was applied yielding approximately a +0.5 gain in BLEU.

2.4.3 ALIBABA (Deng et al., 2018)

Alibaba systems are based on the Transformer model architecture, with the most recent features from the academic research integrated, such as weighted Transformer, Transformer with relative position attention, etc. The system also employs most techniques that have been proven effective during the past WMT years, such as BPE-based subword, back translation, fine-tuning based on selected data, model ensembling and reranking, at industrial scale. For some morphologically-rich languages, linguistic knowledge is also incorporated into the neural network.

2.4.4 CUNI-KOCMI (Kocmi et al., 2018)

The CUNI-KOCMI submission focuses on the low-resource language neural machine translation (NMT). The final submission uses a method of transfer learning: the model is pretrained on a related high-resource language (here Finnish) first, followed by a child low-resource language (Estonian) without any change in hyperparameters. Averaging and backtranslation are also experimented with.

Team	Institution
AALTO	Aalto University (Grönroos et al., 2018)
AFRL	Air Force Research Laboratory (Gwinnup et al., 2018)
ALIBABA	Alibaba Group (Deng et al., 2018)
CUNI-KOCMI	Charles University (Kocmi et al., 2018)
CUNI-TRANSFORMER	Charles University (Popel, 2018)
FACEBOOK-FAIR ⋆	Facebook AI Research (Edunov et al., 2018)
GTCOM	Global Tone Communication Technology (Bei et al., 2018)
HY	University of Helsinki (Raganato et al., 2018)
JHU	Johns Hopkins University (Koehn et al., 2018a)
JUCBNMT	Jadavpur University (Mahata et al., 2018)
KIT	Karlsruhe Institute of Technology (Pham et al., 2018)
LI-MUZE	Li Muze (no associated paper)
LMU-NMT	LMU Munich (Huck et al., 2018)
LMU-UNSUP	LMU Munich (Stojanovski et al., 2018)
MICROSOFT-MARIAN	Microsoft (Junczys-Dowmunt, 2018)
MLLP-UPV	MLLP, Technical University of Valencia (Iranzo-Sánchez et al., 2018)
MMT-PRODUCTION	ModernMT, MMT s.r.l. (no associated paper)
NEUROTOLGE.EE	University of Tartu (Tars and Fishel, 2018)
NICT	National Institute of Information and Communications Technology (Marie et al., 2018)
NIUTRANS	Northeastern University / NiuTrans Co., Ltd. (Wang et al., 2018b)
NJUNMT	NLP Group, Nanjing University (no associated paper)
NTT	NTT Corporation (Morishita et al., 2018)
PARFDA	Boğaziçi University (Biçici, 2018)
PROMT	PROMT LLC (Molchanov, 2018)
RWTH	RWTH Aachen (Schamper et al., 2018)
RWTH-UNSUPER	RWTH Aachen (Graça et al., 2018)
TALP-UPC	TALP, Technical University of Catalonia (Casas et al., 2018)
TENCENT	Tencent (Wang et al., 2018a)
TILDE	Tilde (Pinnis et al., 2018)
UBIQUS	Ubiqus (no associated paper)
UCAM	University of Cambridge (Stahlberg et al., 2018)
UEDIN	University of Edinburgh (Haddow et al., 2018)
UMD	University of Maryland (Xu and Carpuat, 2018)
UNISOUND	Unisound (no associated paper)
UNSUPTARTU	University of Tartu (Del et al., 2018)

Table 2: 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. "⋆" indicates invited participation with a late submission, where the team is not considered a regular participant.

2.4.5 CUNI-TRANSFORMER (Popel, 2018)

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

2.4.6 FACEBOOK-FAIR ⋆ (Edunov et al., 2018)

FACEBOOK-FAIR is an ensemble of six self-attentional models with back-translation data according to Edunov et al. (2018). Synthetic sources are sampled instead of beam search, oversampling the real bitext at a rate of 16, i.e., each bitext is sampled 16 times more often per epoch than the back-translated data. At inference time, translations which are copies of the source are filtered out, replacing them with the output of a very small news-commentary only trained model.

The system FACEBOOK-FAIR has been submitted anonymously as ONLINE-Z and approval for disclosing the authors' identity has only been granted after the final results had become available. Due to the non-standard way of submission, the system is not considered a regular participant, but an invited/late submission and marked with "⋆" throughout the paper.

2.4.7 GTCOM (Bei et al., 2018)

GTCOM-PRIMARY is based on the Transformer "base" model architecture using Marian toolkit, and it also applies some methods that have been proven effective in NMT systems, such as BPE, back-translation, right-to-left reranking and ensembling decoding. In this experiment, right-to-left reranking does not help. Another focus is given to data filtering through rules, translation model and language model including parallel data and monolingual data. The language model is based the Transformer architecture as well. The final system is trained with four different seeds and mixed data.

2.4.8 HY (Raganato et al., 2018; Hurskainen and Tiedemann, 2017)

The University of Helsinki (HY) submitted four systems.

HY-AH (Raganato et al., 2018; Hurskainen and Tiedemann, 2017) is a rule-based machine translation system, relying on a rule-based dependency parser for English, a hand-crafted translation lexicon (based on dictionary data extracted from parallel corpora by word alignment), various types of transfer rules, and a morphological generator for Finnish.

HY-NMT (Raganato et al., 2018) submissions are based on the Transformer "base" model, trained with all the parallel data provided by the shared task plus back-translations, with a shared vocabulary between source and target language and a domain label for each source sentence. For the multilingual sub-track synthetic data for English→Estonian and Estonian→English was also used. Ultimately, a single model for all language pairs was trained and then fine-tuned for each language pair.

HY-NMT-2STEP (Raganato et al., 2018) is a Transformer model trained on interleaved lemmas and morphological tags on the Finnish side. Morphological categories (number, tense etc.) have separate tags, and a tag is only added if the value of the category differs from the default value (in the same way that languages have morphemes only for marked values of morphological categories). The final translation is deterministically generated from the sequence of lemmas and morphological tags which the model outputs.

HY-SMT (Tiedemann et al., 2016) is the Helsinki SMT system submitted at WMT 2016 (the constrained-basic+back-translated version). The system was not retrained and it may thus suffer from poor lexical coverage on recent test data. The main motivation for including this baseline was to have a statistical machine translation (SMT) submission for the Finnish morphology test suite (Burlot et al., 2018).

2.4.9 JHU (Koehn et al., 2018a)

The JHU systems are the result of two relatively independent efforts on German–English language directions and Russian–English, using the Marian and Sockeye (Hieber et al., 2017) neural machine translation toolkits, respectively. The novel contributions are iterative back-translation (for German) and fine-tuning on test sets from prior years (for both languages).

2.4.10 JUCBNMT (Mahata et al., 2018)

JUCBNMT is an encoder-decoder sequence-to-sequence NMT model with character level encoding. The submission uses preprocessing like tokenization, truecasing and corpus cleaning. Both encoder and decoder use a single LSTM layer each. The batch size was set to 128, number of epochs was set to 100, activation function was softmax, optimizer chosen was RMSprop and the loss function used was categorical cross-entropy. Learning rate was set to 0.001.

2.4.11 KIT (Pham et al., 2018)

The KIT submission is the NMT Transformer architecture, enhanced in model depth. Techniques for reducing memory consumption (recalculating intermediate results at layers instead of caching them) allowed 4 times larger model to fit on one GPU and improve the performance by 1.2 BLEU points.

Sentences selection from the new ParaCrawl improved the effectiveness of the corpus by 0.5 BLEU points, with an overall increase of 0.8 BLEU compared to the baseline of not using ParaCrawl.

2.4.12 LI-MUZE

LI-MUZE is an ensemble of 4 averaged Transformer models with one right-to-left and one target-to-source averaged Transformer model, the configuration of all the models is the same as the Transformer big-model, trained on the official training data with 4.5M back-translated data from the monolingual news of 2016 and 2017 data. The English vocabulary size is 36K BPE subwords. Chinese is tokenized by Chinese characters and the vocabulary size is 10K.

2.4.13 LMU-NMT (Huck et al., 2018)

For the WMT18 news translation shared task, LMU Munich (Huck et al., 2018) has trained basic shallow attentional encoder-decoder systems (Bahdanau et al., 2014) with the Nematus toolkit (Sennrich et al., 2017), like last year (Huck et al., 2017a). LMU has participated with these NMT systems for the English–German language pair in both translation directions. The training data is a concatenation of Europarl, News Commentary, Common Crawl, and some synthetic data in the form of backtranslated monolingual news texts. The 2017 monolingual News Crawl is not employed, nor are the parallel Rapid and ParaCrawl corpora. The German data is preprocessed with a linguistically informed word segmentation technique (Huck et al., 2017b). By using a linguistically more sound word segmentation, advantages over plain BPE segmentation are expected in three important aspects: vocabulary reduction, reduction of data sparsity, and open vocabulary translation. The NMT system can learn linguistic word formation processes from the segmented data. In the English→German translation direction, LMU furthermore conducted fine-tuning towards the domain of news articles (Huck et al., 2017a) and reranked the n-best list with a right-to-left neural model (Liu et al., 2016) which is trained for reverse word order (Freitag et al., 2013).

2.4.14 LMU-UNSUP (Stojanovski et al., 2018)

For the unsupervised track of the WMT18 news translation task, LMU Munich submitted the LMU-UNSUP system (Stojanovski et al., 2018) which is a neural translation model trained without any access to parallel data. The model is trained with ∼4M German and English sentences each, which are sampled from NewsCrawl articles from 2007 to 2017. Bilingual word embeddings trained in an unsupervised manner (Conneau et al., 2017) were used to translate the monolingual data by doing word-by-word translation and this synthetically created parallel data is used in the training as well. The same model is used to do both German→English and English→German translation. The model is based on Lample et al. (2018) and it uses denoising and on-the-fly backtranslation. Additionally the model uses the word-by-word translated data in the initial training stages to jump-start the training and disables the denoising component as the last training step for further improvements. The NMT embeddings are initialized with embeddings obtained from `fasttext` trained jointly on German and English monolingual BPE-level data.

2.4.15 MICROSOFT-MARIAN (Junczys-Dowmunt, 2018)

MICROSOFT-MARIAN is the Transformer-big model implemented in Marian with an updated version of Edinburgh's training scheme for WMT2017, following current common practices: truecasing and tokenization using Moses scripts, BPE subwords, backtranslation (using a shallow model), ensembling of four left-to-right deep models and reranking of 12-best list with an en-

semble of four right-to-left models.

The novelties are primarily in new data filtering (dual conditional cross-entropy filtering) and sentence weighting methods.

2.4.16 MLLP-UPV (Iranzo-Sánchez et al., 2018)

MLLP-UPV is an ensemble of Transformer architecture-based neural machine translation systems. To train the system under "constrained" conditions, the provided parallel data was filtered with a scoring technique using character-based language models, and was augmented based on synthetic source sentences generated from the provided monolingual corpora.

The ensemble consists of 4 independent training runs of the Transformer "base" model, trained with 10M filtered sentences (including from ParaCrawl) and 20M backtranslated sentences from NewsCrawl2017.

2.4.17 MMT-PRODUCTION

MMT-PRODUCTION is the machine translation system offered by MMT s.r.l. (www.modernmt.eu) as of July 2018. It is a Transformer-based neural MT system trained on public and proprietary data, containing about 100M sentence pairs and about 1.5G English words. It exploits a single model of 'transformer-big' size, and a single pass-decoding; texts are processed using internal tools.

2.4.18 NEUROTOLGE.EE (Tars and Fishel, 2018)

NEUROTOLGE.EE is a multi-domain NMT system that treats text domain as language and applies the zero-shot multi-lingual approach to multiple domains in the training corpus. For WMT18, text domains were replaced with unsupervised clustering into 16 clusters using FastText's sentence embeddings. During translation the input segment is classified using its sentence embedding and translated as the corresponding cluster/domain.

2.4.19 NICT (Marie et al., 2018)

NICT NMT systems were trained with the Transformer architecture using the provided parallel data enlarged with a large quantity of back-translated monolingual data generated with a new incremental training framework. The primary submissions to the task are the result of a simple combination between NICT SMT and NMT systems.

2.4.20 NIUTRANS (Wang et al., 2018b)

NIUTRANS baseline systems are based on the Transformer architecture with the "base" model, equipped with checkpoint averaging and back-translation techniques. NIUTRANS further improves the translation performance by 2.28–3.83 BLEU points from four aspects of model variations (larger inner-hidden-size in FFN, using ReLU and attention dropout, Swish activation function, relative positional representation), diverse ensemble decoding (ensemble decoding with up to 15 models, generated by different strategies), reranking (up to 14 features for reranking), and post-processing (aimed at consistent translation of proper nouns, especially English literals in Chinese sentences).

2.4.21 NJUNMT

The NJUNMT-PRIVATE is most likely the system developed by Natural Language Processing Group of Nanjing University based on high-level API of TensorFlow, https://github.com/zhaocq-nlp/NJUNMT-tf. Further details on training are not available.

2.4.22 NTT (Morishita et al., 2018)

NTT combine Transformer "big" model, corpus cleaning technique for provided and synthetic parallel corpora, and right-to-left n-best re-ranking techniques. Through their experiments, NTT found filtering of noisy training sentences and right-to-left re-ranking as the keys to better accuracy.

2.4.23 PARFDA (Biçici, 2018)

PARFDA selects a subset of the training and LM data to build task-specific SMT models. PARFDA uses phrase-based Moses and all constrained available resources provided by WMT18. The datasets are available at https://github.com/bicici/parfdaWMT2018.

2.4.24 PROMT (Molchanov, 2018)

PROMT submitted three systems: PROMT-HYB-MARIAN, PROMT-HYB-OPENNMT and PROMT-RULE-BASED.

PROMT-HYB-MARIAN is an ensemble of 5 transformer models trained on WMT data and in-house news data.

PROMT-HYB-OPENNMT is a hybrid system based on PROMT Rule-based engine and an NMT

post-editing (PE) engine. The NMT PE component is a sequence-to-sequence model with attention and deep biRNN encoder trained with OpenNMT toolkit.

PROMT-RULE-BASED is a rule-based system, without any specific training or tuning.

2.4.25 RWTH (Schamper et al., 2018)

All systems submitted by RWTH Aachen for German to English are based on the Transformer architecture implemented in Sockeye. The final RWTH system has been an ensemble of three Transformer models, where each individual model had been already very strong. The strength of the RWTH systems is probably due to the following four key factors: (a) Using the Transformer architecture. (b) Rather large models and large batch size which was made possible due to synchronous training on 4 GPUs and roughly 8 days of training. (Details: num-embed: 1024; num-layers: 6; attention-heads: 16; transformer-feed-forward-num-hidden: 4096; transformer-model-size: 1024, no weight-tying. In sum, this results in 291M trainable parameters.) (c) Careful experiments on data conditions: E.g. oversampling of parallel data, LM driven filtering of ParaCrawl (retained 50%), testing different amounts of BPE merge operations. (d) Fine-tuning on old testsets (newstest2008-newstest2014).

RWTH English→Turkish system is based on 6-layer encoder-decoder Transformer architecture. Since the task has low resources, dropout with the rate of 0.3 to all applicable layers was used. Even though the two languages are not much related, joint BPE and weight tying helped a lot as part of regularization. For the final submission, RWTH used augmented training data with 1M-sentence back-translations and ensembled four models with different random seeds.

2.4.26 RWTH-UNSUPER (Graça et al., 2018)

The RWTH-UNSUPER unsupervised NMT system is built based on recent works by Lample et al. (2018) and Artetxe et al. (2018). RWTH-UNSUPER best performing systems follow the batch optimization strategy and are initialized with cross-lingual embeddings. Furthermore, RWTH-UNSUPER found that sharing a vocabulary performs better than having separate ones. Freezing embeddings hurts performance and it was found best to initialize embeddings with pre-trained ones and train them as usual.

2.4.27 TALP-UPC (Casas et al., 2018)

TALP-UPC is the Transformer "base" model trained with the Tensor2Tensor implementation (Vaswani et al., 2018) and wordpieces vocabulary. The training corpus is multilingual (concatenating Finnish–English and Estonian–English) and includes ParaCrawl with garbage cleaned up via `langdetect`.

2.4.28 TENCENT (Wang et al., 2018a)

TENCENT-ENSEMBLE (called TenTrans) is an improved NMT system on Transformer based on self-attention mechanism. In addition to the basic settings of Transformer training, TENCENT-ENSEMBLE uses multi-model fusion techniques, multiple features reranking, different segmentation models and joint learning. Additionally, data selection strategies were adopted to fine-tune the trained system, achieving a stable performance improvement.

An additional system paper (Hu et al., 2018) describes a non-primary submission.

2.4.29 TILDE (Pinnis et al., 2018)

TILDE submitted four systems: TILDE-C-NMT, TILDE-C-NMT-COMB, TILDE-C-NMT-2BT and TILDE-NC-NMT.

TILDE-C-NMT are constrained English-Estonian and Estonian-English NMT systems that were deployed as ensembles of averaged factored data Transformer models. The models were trained using filtered parallel data and back-translated data in a 1-to-1 proportion. The parallel data were supplemented with synthetic data (generated from the same parallel data) that contain unknown token identifiers in order to acquire models that are more robust to unknown phenomena.

TILDE-C-NMT-COMB is a constrained Estonian-English NMT system that is a system combination of multiple constrained factored data NMT systems.

TILDE-C-NMT-2BT systems were trained using Sockeye and Transformer models. Before training the initial systems, parallel data were cleaned using the `parallel-corpora-tools`. Before back-translation, monolingual data were also filtered. After back-translation, the resulting synthetic corpora were filtered again. Intermediate

systems were trained with the first batch of parallel+synthetic data. The back-translation and filtering process was performed a second time with additional monolingual data to train the final systems with parallel and two sets of synthetic data.

TILDE-NC-NMT are unconstrained English→Estonian and Estonian→English NMT systems that were deployed as averaged Transformer models. These models were also trained using back-translated data similarly to the constrained systems, however, the data, taking into account its relatively large size, was not factored.

2.4.30 UBIQUS

UBIQUS-NMT is a Transformer "base" model trained and run with the OpenNMT implementation. It uses back-translation according to Sennrich et al. (2016a) and it does not include ParaCrawl. Subwords are generated with Sentence Piece.[9]

2.4.31 UCAM (Stahlberg et al., 2018)

UCAM is a generalization of previous work (de Gispert et al., 2017) to multiple architectures. It is a system combination of two Transformer-like models, a recurrent model, a convolutional model, and a phrase-based SMT system. The output is probably dominated by the Transformer, and to some extend by the SMT system.

2.4.32 UEDIN (Haddow et al., 2018)

For Estonian↔English and Finnish↔English, the UEDIN systems are an ensemble of four left-to-right systems, reranked with four right-to-left systems, built using Marian. Each ensemble consists of two Transformers and two deep RNNs. The RNNs use the UEDIN multi-head / multi-hop variant. All available parallel data were used, plus back-translated data from 2017 (for into-English) and 2014-2017 (for out-of-English). The natural parallel data was generally over-sampled to give an equal mix of parallel and synthetic data. For English↔Estonian, UEDIN selected 30% of ParaCrawl based on translation model perplexity for a model built on the rest of the data.

The UEDIN systems for other language pairs use an ensemble of four deep RNN left-to-right systems, reranked with 4 deep RNN right-to-left systems. The RNN models use the UEDIN multi-head / multi-hop attention variant. All the provided parallel data (including ParaCrawl) were used, applying `langid` filtering to remove some incorrect sentence pairs. Synthetic data were also used, created by back-translating the 2017 English news crawl, and the 2017 and 2016 Czech news crawls. For Czech→English, the synthetic data was oversampled 2x.

2.4.33 UMD (Xu and Carpuat, 2018)

The UMD best system is an ensemble of three 6-layer left-to-right Transformer models reranked with target-to-source and left-to-right models. Each Transformer model is trained with a 2:1 mixture of parallel and backtranslated monolingual data. For parallel data, duplicates are removed and "bad" sentence pairs filtered out. Monolingual data is sub-sampled from news 2017 (English) and news 2011 (Chinese). Subwords (BPE) are used for both English and Chinese sentences.

2.4.34 UNISOUND

The UNISOUND systems are probably developed by the Unisound company (`www.unisound.com`). No further information is available.

2.4.35 UNSUPTARTU (Del et al., 2018)

UNSUPTARTU is an unsupervised MT system using n-gram embedding cross-lingual mapping to create a phrase table. An RNN LM is used in decoding.

2.5 Submission Summary

Next we summarize the general trends in the systems submitted to the translation task and its subtracks.

The dominating majority of the submissions (29 systems) are based on the Transformer approach (Vaswani et al., 2017), with a varying number of encoder/decoder layers and other details. Four more systems use the basic attentional encoder-decoder approach (Bahdanau et al., 2014), three are phrase-based SMT systems and two are rule-based. Several submissions use ensembles of components with different approaches.

Most systems report using back-translated data, some of them filtering the synthetic data and some using a fixed sampling rate between the real and synthetic data.

As far as subwords go, two widely used options are byte-pair encoding (Sennrich et al., 2016b) and sentencepiece (Kudo and Richardson, 2018).

[9] `https://github.com/google/sentencepiece`

Some submissions use linguistically motivated segmentation, especially for the highly agglutinative Finnish.

There were 3 submissions to the multilingual sub-track, all three applying multilingual transfer learning and training systems to translate from English into Finnish and Estonian simultaneously.

Unsupervised MT also had 3 submissions, of which two applied their systems to German to/from English and the third was done for Estonian-to-English translation. The two German-English-German systems use the neural MT method of Lample et al. (2018) with small modifications, and the Estonian-English system used a phrase-based statistical unsupervised approach from the same article.

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. 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.

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 two 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., 2016a), 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). Last year, we thus employed DA for evaluation of systems taking part in the news task and do so again this year. Where possible, we collect DA judgments via the crowd-sourcing platform, Amazon's Mechanical Turk, and as in previous year's we ask participating teams to provide manual evaluation of system outputs via Appraise. Researcher involvement was needed particularly for translations into Czech, German, Estonian, Finnish and Turkish.

Human assessors are asked to rate a given translation by how adequately it expresses the meaning of the corresponding reference translation (i.e. no bilingual speakers are needed) on an analogue scale, which corresponds to an underlying absolute 0–100 rating scale. Since DA involves evaluation of a single translation per screen, this allows the sentence length restriction usually applied during manual evaluation to be removed for both researchers and crowd-sourced workers.[10] Figure 3 shows one DA screen as completed by researchers on Appraise, while Figure 4 provides a screenshot of DA shown to crowd-sourced workers on Amazon's Mechanical Turk.

The 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. Appraise users were allowed to pause their annotation at any time, Amazon interface did not allow any pauses. More details of composition of HITs are given in Section 3.3 below.

3.1 Evaluation Campaign Overview

In terms of the News translation task manual evaluation, a total of 584 individual researcher accounts were involved, and 915 turker accounts.[11] Researchers in the manual evaluation came from 33 different research groups and contributed judgments of 118,705 translations, while 225,900 translation assessment scores were submitted in total by the crowd.[12]

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, we are often able to take advantage of this and use a single assessment for multiple systems. Similar to last year's evaluation, we only combine human assessments in this way if the string of text belonging to multiple systems is exactly identical. For example, even small dif-

[10] The maximum sentence length with RR was 30 in WMT16.

[11] Numbers do not include the 1,533 workers on Mechanical Turk and 7 on Appraise who did not pass quality control.

[12] Numbers include quality control items for workers who passed quality control but omit the additional 347,700 assessments collected on Mechanical Turk where a worker did not pass quality control and equivalent 1,466 judgments for the small number of Appraise workers who did not meet the quality control threshold. A 40% pass rate for quality control is typical of DA evaluations on Mechanical Turk.

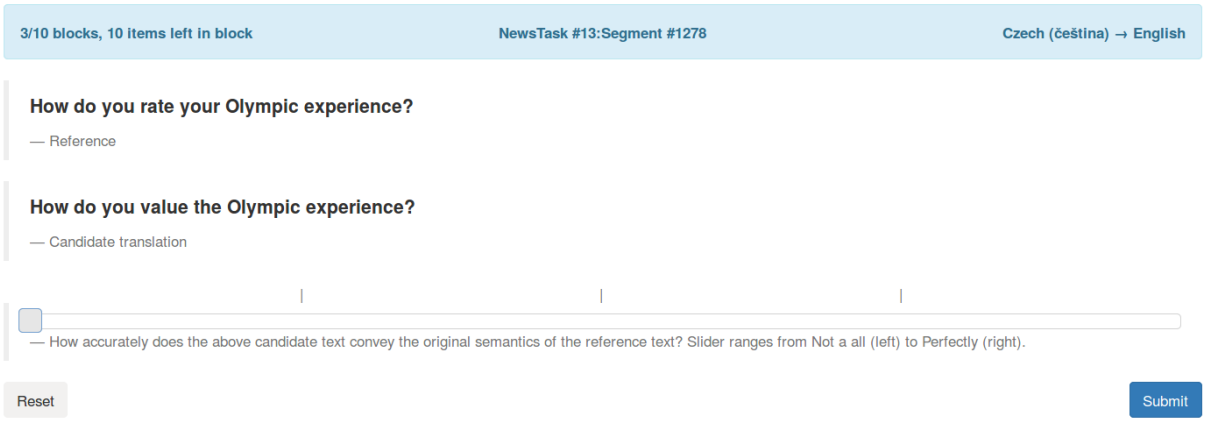

Figure 3: Screen shot of Direct Assessment in the Appraise interface used in the human evaluation campaign. The annotator is presented with a reference translation and a single system output randomly selected from competing systems (anonymized), and is asked to rate the translation on a sliding scale.

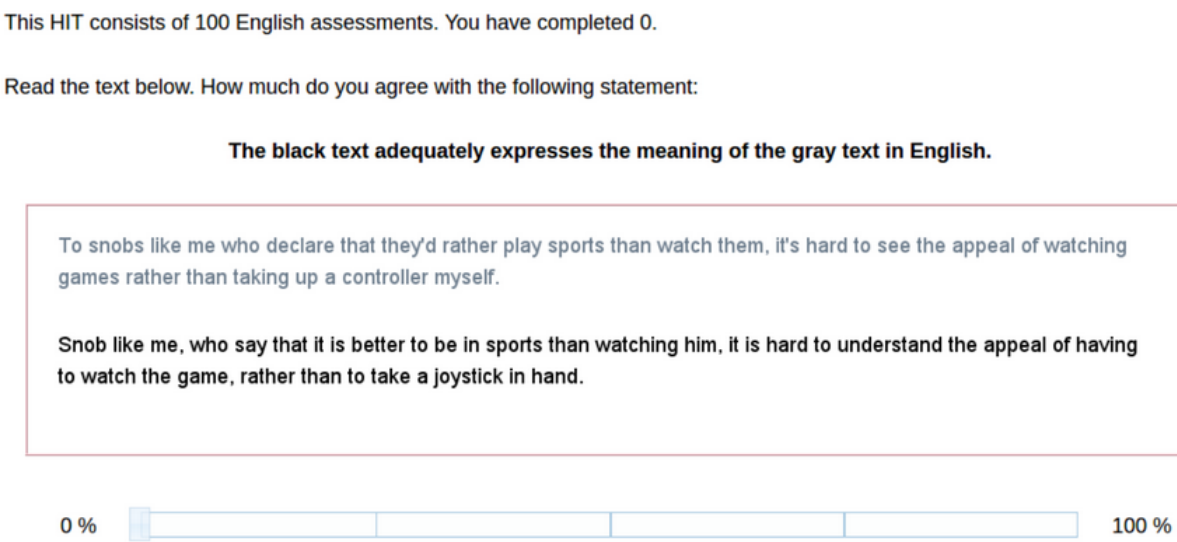

Figure 4: Screen shot of Direct Assessment as carried out by workers on Mechanical Turk.

ferences in punctuation disqualify combination of similar system outputs, and this is due to a general lack of evidence about what kinds of minor differences may or may not impact human evaluation.

Table 3 shows the numbers of segments for which distinct MT systems participating in the News Translation Task produced identical outputs. The biggest saving in terms of exact duplicate translations, being produced by multiple systems, was for German to English, where a 17.4% saving of resources by combining identical outputs before human evaluation.

3.2 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[13] (Federmann, 2012) tool for our data collection, in addition to Amazon Mechanical Turk.[14] Table 4 shows total numbers of human assessments collected in WMT18 contributing to final scores for systems.[15]

The effort that goes into the manual evaluation campaign each year is impressive, and we

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

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

[15] Appraise ran evaluation of $150-1=149$ systems due to a single tr-en system having been omitted in the initial human evaluation run. The 95 crowd-sourced systems includes all into-English language pair (including the tr-en missing system), en-ru and en-zh systems.

Language Pair	Systems	Segments	Total Segments	Distinct Segments	Saving (%) WMT18	Saving (%) WMT17
Chinese→English	14	3,981	55,734	49,767	**10.7**	3.9
Czech→English	5	2,983	14,915	13,987	**6.2**	4.3
German→English	16	2,998	47,968	39,627	**17.4**	10.7
Estonian→English	14	2,000	28,000	25,612	**8.5**	–
Finnish→English	9	3,000	27,000	25,233	**6.5**	1.4
Russian→English	8	3,000	24,000	21,966	**8.5**	5.8
Turkish→English	6	3,000	18,000	17,000	**5.6**	4.6
English→Chinese	14	3,981	55,734	48,022	**13.8**	1.7
English→Czech	5	2,983	14,915	13,982	**6.3**	10.2
English→German	16	2,998	47,968	39,963	**16.7**	12.8
English→Estonian	14	2,000	28,000	25,837	**7.7**	–
English→Finnish	12	3,000	36,000	32,749	**9.0**	3.7
English→Russian	9	3,000	27,000	24,594	**8.9**	4.5
English→Turkish	8	3,000	24,000	21,880	**8.8**	2.1

Table 3: Total segments prior to sampling for manual evaluation and savings made by combining outputs produced by different systems that were identical.

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.

3.3 Crowd Quality Control

This year, two distinct HIT structures were run in the overall evaluation campaign, the standard DA set-up was employed for Mechanical Turk and a portion of the Appraise evaluation, while an additional HIT structure was used for the remaining part of the Appraise evaluation. Below we firstly describe the standard DA HIT structure and quality control mechanism before describing the additional version used for part of the Appraise evaluation. In both set-ups, translations are arranged in sets of 100-translation HITs to provide control over assignment and positioning of quality control items to human annotators.

3.3.1 Standard DA HIT Structure

In the standard DA HIT structure, three kinds of quality control translation pairs are employed as described in Table 5: 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 control items, HITs are constructed using translations sampled from the entire set of outputs for a given language pair.

Language Pair	Systems	Comps	Comps/Sys	Assessments	Assess/Sys
Chinese→English	14	–	–	32,919	2,351.4
Czech→English	5	–	–	12,209	2,441.8
German→English	16	–	–	48,469	3,029.3
Estonian→English	14	–	–	28,868	2,062.0
Finnish→English	9	–	–	18,868	2,096.4
Russian→English	8	–	–	17,711	2,213.9
Turkish→English	6	–	–	29,784	4,964.0
English→Chinese	14	–	–	32,411	2,315.1
English→Czech	5	–	–	10,080	2,016.0
English→German	16	–	–	13,754	859.6
English→Estonian	14	–	–	15,800	1,128.6
English→Finnish	12	–	–	9,995	832.9
English→Russian	9	–	–	27,977	3,108.6
English→Turkish	8	–	–	3,644	455.5
Total Researcher	149	–	–	101,189	679.1
Total Crowd	95	–	–	201,300	2,118.9
Total 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 4: Amount of data collected in the WMT18 manual evaluation campaign (assessments after removal of quality control items and "de-collapsing" *multi-system outputs*). The final seven rows report summary information from previous years of the workshop.

3.3.2 Alternate DA HIT Structure

The alternate DA HIT structure employed by Appraise this year for a subset of researcher HITs is shown in Table 6. This set-up reduces the number of quality control items in a HIT and is therefore more efficient (12% overhead) by omitting repeat pairs and good reference pairs. This comes at the cost of a reduced ability to analyze the quality of data provided by human annotators.

In addition for this set-up, an additional constraint (not originally applied in standard DA) was imposed. As much as possible within a 100-translation HIT the HIT included the output of all participating systems for each source input. This constraint has the advantage of producing assessments from the same human assessor for translations of the same source input but is not ideal in terms of the original aim of DA – to as much as possible produce absolute scores for translations (as opposed to relative ones) – because it positions assessment of competing translations in close proximity within a HIT and judges may attempt to remember their judgment for a different candidate translation of a given input sentence.

3.3.3 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

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

Bad Reference Pairs:	Original System output (12)	A degraded version of it (12).

Table 6: Alternate DA HIT structure used for a portion of researchers in Appraise data collection, where quality control translation pairs hidden within 100-translation HITs, numbers of items are provided in parentheses in adapted version of DA used for a subset of researchers HITs.

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 ⌋

3.4 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.

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.

This year's assessment includes the first large-scale DA evaluation where quality control items were applied to assessments of a known-reliable group, comprised of the portion of researchers who completed HITs on Appraise with the original DA HIT structure. Although this group should be considered highly reliable compared to Mechanical Turk for example, we must however keep in mind that a small part of this group are in fact hired to complete assessments and their reliability could vary more than what would be expected of volunteer researchers.

Table 7 shows the number of workers in the crowd-sourced and researcher groups who met our filtering requirement by showing a significantly lower score for bad reference items compared to corresponding MT outputs, and the proportion of those who simultaneously showed no significant difference in scores they gave to pairs of identical translations.

The main observation to be taken from Table 7 is the difference in proportions of human assessors on Mechanical Turk versus researchers who passed the quality filtering criteria for DA, by scoring degraded translations significantly lower than the original MT output counterparts, as 37% of Mechanical Turk workers were deemed reliable compared to 93% of evaluators in the researcher group. This low rate of workers passing quality filtering is in line with past DA evalu-

ations, and the high proportion of annotators passing quality control is expected of a mostly known-reliable group. For crowd-sourced workers, consistent with past DA evaluations, Table 7 shows a substantially higher number of low quality workers encountered for evaluation of languages other than English on Mechanical Turk. For example, in the case of Russian and Chinese only a respective 22% and 10% of workers were considered reliable enough to include their assessments in the evaluation, compared to around 42% on average for English evaluations.

When we examine repeat assessments of the same translation, both filtered groups show similar levels of reliability with 96% of filtered Mechanical Turk workers and 95% of researchers showing no significant difference in scores for repeat assessment of the same translation. The idea is that the repeated input should receive a very similar score. Assuming that annotators do not remember their previous assessment for the repeated sentence, the "Exact Rep." corresponds to intra-annotator agreement and it reaches very high scores.[16]

Within the researcher group, although assessors have high levels of reliability overall, reliability in this respect varies quite a bit for different languages. For example, only 75% of assessors in the researcher group completing assessments for Estonian showed no significant difference for repeat assessment of the same translation, and 87% for Turkish, both lower levels of reliability than usually encountered on Mechanical Turk even though the research group is expected to be more reliable than crowd-sourced workers. However, on closer inspection, the number of human assessors who took part in the Turkish and Estonian evaluations is small and the seemingly large difference in percentages in fact correspond to as few as three individuals.

3.5 Producing the Human Ranking

All research and crowd data that passed quality control were combined to produce the overall shared task results. 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 both researchers and crowd. Average standardized scores for individual segments belonging to a given system are then computed, before the final overall DA score for that system is computed as the average of its segment scores (Ave z in Table 8). Results are also reported for average scores for systems, computed in the same way but without any score standardization applied (Ave % in Table 8).

Table 8 includes final DA scores for all systems participating in WMT18 News Translation Task. 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.

Note that for English→German, the system FACEBOOK-FAIR is not considered a regular participant, but an invited/late submission, see Section 2.4.6.

Appendix A shows the underlying head-to-head significance test results for all pairs of systems.

3.6 Source-based Direct Assessment

A secondary bilingual manual evaluation was carried out involving an adaptation of the standard monolingual DA evaluation in which the source language input segment was used in place of the reference. Figure 5 provides a screenshot of this evaluation as implemented in Appraise, which we refer to as *source-based DA*. In this set-up system outputs are evaluated by bilinguals who have access to the source language input segment only and no reference translation. The main motivation for doing so was to free up reference translations to allow them to be used instead as a "human system" in the evaluation. By structuring the evaluation as a bilingual task it allows a human system to be manually evaluated under exactly the same conditions as all other systems thus providing an estimate of human performance.[17]

The aim of source-based DA is to produce accurate rankings for systems as well as the human system to allow direct comparison of system and human performance, motivated by recent indications that Machine Translation quality may in some cases be approaching human performance (Wu et al., 2016; Hassan et al., 2018). For source-based DA, annotators will ideally be bilingual, i.e.

[16] Repeat items are separated by a minimum of 40 intervening assessments to reduce the likelihood of annotators simply remembering previous scores for repeat assessment of translations.

[17] An alternate method is to keep DA monolingual but to employ secondary reference translations. No secondary reference translations were available for the test set, however.

		All	(A) Sig. Diff. Bad Ref.	(B) (A) & No Sig. Diff. Exact Rep.
Mechanical Turk Crowd	Czech→English	169	74 (44%)	70 (95%)
	German→English	514	227 (44%)	216 (95%)
	Estonian→English	397	157 (40%)	150 (96%)
	Finnish→English	238	102 (43%)	99 (97%)
	Russian→English	203	96 (47%)	93 (97%)
	Turkish→English	480	172 (36%)	166 (97%)
	Chinese→English	401	153 (38%)	148 (97%)
	English→Russian	209	47 (22%)	45 (96%)
	English→Chinese	406	39 (10%)	37 (95%)
	Crowd	**2,477**	**915 (37%)**	**880 (96%)**
Researcher	German→English	41	39 (95%)	37 (95%)
	Estonian→English	16	13 (81%)	13 (100%)
	Finnish→English	3	3 (100%)	3 (100%)
	Russian→English	8	8 (100%)	8 (100%)
	Turkish→English	7	7 (100%)	7 (100%)
	Chinese→English	4	3 (75%)	3 (100%)
	English→Czech	17	17 (100%)	17 (100%)
	English→German	48	47 (98%)	44 (94%)
	English→Estonian	6	4 (67%)	3 (75%)
	English→Finnish	29	27 (93%)	25 (93%)
	English→Russian	26	25 (96%)	24 (96%)
	English→Turkish	17	15 (88%)	13 (87%)
	English→Chinese	34	31 (91%)	30 (97%)
	Researcher	**256**	**239 (93%)**	**227 (95%)**
Researcher$_{alt}$	Czech→English	32	30 (94%)	—
	German→English	41	39 (95%)	—
	Estonian→English	12	12 (100%)	—
	Finnish→English	4	3 (75%)	—
	Russian→English	7	5 (71%)	—
	Turkish→English	3	2 (66%)	—
	Chinese→English	4	4 (100%)	—
	English→Czech	49	49 (100%)	—
	English→German	31	31 (100%)	—
	English→Estonian	83	83 (100%)	—
	English→Finnish	30	30 (100%)	—
	English→Russian	37	36 (97%)	—
	English→Turkish	6	6 (100%)	—
	English→Chinese	23	22 (96%)	—
	Researcher$_{alt}$	**362**	**352 (97%)**	—
	Total WMT18	**3,095**	**1,506 (49%)**	**1,107 (96%)**

Table 7: Number of unique workers, (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. *Researcher* denotes the portion of the evaluation carried out with the standard DA HIT structure, while *Researcher$_{alt}$* denotes the remaining part that employed the altered HIT structure in which some quality control items are omitted.

Chinese→English

	Ave. %	Ave. z	System
1	78.8	0.140	NIUTRANS
	77.7	0.111	ONLINE-B
	77.9	0.109	UCAM
	78.0	0.108	UNISOUND-A
	77.5	0.099	TENCENT-ENSEMBLE
	77.5	0.094	UNISOUND-B
	77.9	0.091	LI-MUZE
	77.0	0.089	NICT
	76.7	0.078	UMD
10	75.0	−0.005	ONLINE-Y
	74.5	−0.017	UEDIN
12	73.6	−0.061	ONLINE-A
13	65.9	−0.327	ONLINE-G
14	64.4	−0.377	ONLINE-F

English→Chinese

	Ave. %	Ave. z	System
1	80.7	0.219	TENCENT-ENSEMBLE
	80.3	0.206	UNISOUND
	80.5	0.199	GTCOM-PRIMARY
	79.7	0.185	ALIBABA-ENS-RERANK
	79.2	0.173	ALIBABA-GENERAL-A
	79.5	0.166	ONLINE-B
	79.0	0.165	ALIBABA-GENERAL-B
8	78.1	0.094	UMD
	77.5	0.082	NICT
	77.1	0.069	ONLINE-Y
	75.5	0.037	ONLINE-A
12	70.7	−0.202	UEDIN
13	63.3	−0.419	ONLINE-F
	63.4	−0.435	ONLINE-G

Czech→English

	Ave. %	Ave. z	System
1	71.8	0.298	CUNI-TRANSFORMER
2	67.9	0.165	UEDIN
3	66.6	0.115	ONLINE-B
4	62.1	−0.023	ONLINE-A
5	57.5	−0.183	ONLINE-G

English→Czech

	Ave. %	Ave. z	System
1	67.2	0.594	CUNI-TRANSFORMER
2	60.6	0.384	UEDIN
3	52.1	0.101	ONLINE-B
4	46.0	−0.115	ONLINE-A
5	42.0	−0.246	ONLINE-G

German→English

	Ave. %	Ave. z	System
1	79.9	0.413	RWTH
	79.4	0.395	UCAM
	78.2	0.359	NTT
	77.3	0.346	ONLINE-B
	77.4	0.321	MLLP-UPV
	77.0	0.317	JHU
	76.9	0.315	UBIQUS-NMT
	76.7	0.310	ONLINE-Y
	75.7	0.268	ONLINE-A
	75.4	0.261	UEDIN
11	72.5	0.162	LMU-NMT
	72.2	0.149	NJUNMT-PRIVATE
13	65.2	−0.074	ONLINE-G
14	58.5	−0.296	ONLINE-F
15	45.4	−0.752	RWTH-UNSUPER
16	42.7	−0.835	LMU-UNSUP

English→German

	Ave. %	Ave. z	System
1	85.5	0.653	FACEBOOK-FAIR ⋆
2	82.2	0.561	ONLINE-B
	81.9	0.551	MICROSOFT-MARIAN
	81.6	0.539	MMT-PRODUCTION
	82.3	0.537	UCAM
	80.2	0.491	NTT
	79.3	0.454	KIT
8	77.7	0.396	ONLINE-Y
	76.7	0.377	JHU
	76.3	0.352	UEDIN
11	71.8	0.213	LMU-NMT
12	67.4	0.060	ONLINE-A
13	53.2	−0.385	ONLINE-F
	53.8	−0.416	ONLINE-G
15	36.7	−0.966	RWTH-UNSUPER
16	32.6	−1.122	LMU-UNSUP

Estonian→English

	Ave. %	Ave. z	System
1	73.3	0.326	TILDE-NC-NMT
2	71.1	0.238	NICT
	69.9	0.215	TILDE-C-NMT
	69.0	0.187	TILDE-C-NMT-2BT
	69.2	0.186	UEDIN
	68.7	0.171	TILDE-C-NMT-COMB
	67.1	0.117	ONLINE-B
	66.4	0.106	HY-NMT
	66.8	0.106	TALP-UPC
10	65.4	0.063	ONLINE-A
	64.0	0.007	CUNI-KOCMI
12	59.4	−0.117	NEUROTOLGE.EE
13	52.7	−0.341	ONLINE-G
14	34.6	−0.950	UNSUPTARTU

English→Estonian

	Ave. %	Ave. z	System
1	64.9	0.549	TILDE-NC-NMT
2	62.1	0.453	NICT
	61.6	0.427	TILDE-C-NMT
	61.2	0.418	TILDE-C-NMT-2BT
5	58.6	0.340	AALTO
	58.6	0.329	HY-NMT
	57.5	0.295	UEDIN
8	55.5	0.216	CUNI-KOCMI
	54.6	0.181	TALP-UPC
10	52.1	0.097	ONLINE-B
11	45.7	−0.132	NEUROTOLGE.EE
12	43.8	−0.195	ONLINE-A
13	37.6	−0.406	ONLINE-G
14	34.3	−0.520	PARFDA

Finnish→English

	Ave. %	Ave. z	System
1	75.2	0.153	NICT
	74.4	0.128	HY-NMT
	74.0	0.103	UEDIN
	72.7	0.083	CUNI-KOCMI
	72.9	0.078	ONLINE-B
	71.9	0.047	TALP-UPC
	71.5	0.045	ONLINE-A
8	66.1	−0.134	ONLINE-G
9	58.9	−0.404	JUCBNMT

English→Finnish

	Ave. %	Ave. z	System
1	64.7	0.521	NICT
	63.1	0.466	HY-NMT
3	59.2	0.324	UEDIN
	58.3	0.271	AALTO
	57.9	0.258	HY-NMT-2STEP
	57.4	0.238	TALP-UPC
	55.9	0.184	CUNI-KOCMI
	56.6	0.183	ONLINE-B
9	45.9	−0.212	ONLINE-A
	45.3	−0.233	ONLINE-G
11	42.7	−0.334	HY-SMT
	41.5	−0.369	HY-AH

Russian→English

	Ave. %	Ave. z	System
1	81.0	0.215	ALIBABA
	80.3	0.192	ONLINE-B
	79.6	0.170	ONLINE-G
4	77.5	0.110	UEDIN
5	76.2	0.034	ONLINE-A
6	74.1	−0.014	AFRL-SYSCOMB
	73.7	−0.027	JHU
8	64.2	−0.398	ONLINE-F

English→Russian

	Ave. %	Ave. z	System
1	72.0	0.352	ALIBABA-ENS
	71.4	0.324	ONLINE-G
3	66.8	0.159	ONLINE-B
	66.0	0.144	UEDIN
	64.9	0.115	PROMT-HYB-MARIAN
6	63.9	0.066	PROMT-HYB-OPENNMT
7	62.2	−0.004	ONLINE-A
8	59.1	−0.075	PROMT-RULE-BASED
9	44.5	−0.580	ONLINE-F

Turkish→English

	Ave. %	Ave. z	System
1	70.2	0.101	ONLINE-G
	69.3	0.077	ONLINE-A
	68.1	0.030	ALIBABA-ENS
	68.0	0.027	ONLINE-B
	67.0	−0.008	UEDIN
	66.0	−0.040	NICT

English→Turkish

	Ave. %	Ave. z	System
1	66.3	0.277	ONLINE-B
	63.6	0.222	UEDIN
	63.5	0.216	ALIBABA-ENS-A
	62.0	0.128	NICT
	60.1	0.111	ALIBABA-ENS-B
	60.1	0.058	ONLINE-G
7	55.0	−0.060	RWTH
8	49.6	−0.254	ONLINE-A

Table 8: Official results of WMT18 News Translation Task. Systems ordered by standardized mean DA score, though systems within a cluster are considered tied. Lines between systems indicate clusters according to Wilcoxon rank-sum test at p-level $p < 0.05$. Systems with gray background indicate use of resources that fall outside the constraints provided for the shared task.

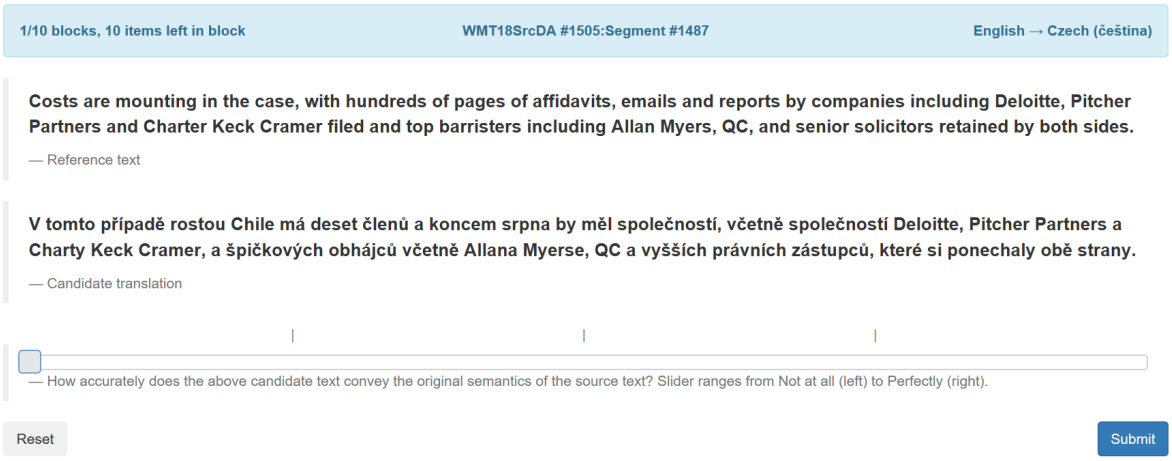

Figure 5: Screen shot of source-based Direct Assessment in the Appraise interface used in the English→Czech pilot campaign. The annotator is presented with a source text and a single system output randomly selected from competing systems (anonymized), and is asked to rate the translation on a sliding scale.

understand the source language sufficiently well, in addition to being native speakers of the target language. However, we did not specifically stipulate in this year's evaluation that human annotators be native speakers of the target language.

We run source-based DA for evaluation of English to Czech translation. This language pair was selected because sufficient annotators were available, helped by the fact that the set of systems participating in this language pair is small. This part of the campaign employs the alternate HIT structure described in Section 3.3.2 with reduced quality control items, i.e. it does not include exact repeats of translations or reference translations for quality control purposes.

A total of 17 annotators worked on the source-based DA pilot. 100% of annotators proved reliable, meaning that they scored bad reference items significantly lower than corresponding MT outputs (see Table 7 part (A) for corresponding reference-based DA percentages). For six candidate systems we collected 2,574 assessments, resulting in an average of 429 annotations per individual system. Enforcing segment overlap during HIT creation resulted in 423 segments for which all six candidate translations have been scored. In total, annotators worked on 438 distinct segments.

Table 9 provides source-based DA scores for all primary English→Czech systems participating in WMT18 News Translation Task as well as the human system comprised of reference translations. 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.

English→Czech

	Ave. %	Ave. z	System
1	84.4	0.667	CUNI-TRANSFORMER
2	79.8	0.521	UEDIN
	78.6	0.483	NEWSTEST2018-REF
4	68.1	0.128	ONLINE-B
5	59.4	−0.178	ONLINE-A
6	54.1	−0.354	ONLINE-G

Table 9: Source-based DA results for English→Czech newstest2018, where systems are ordered by standardized mean DA score, though systems within a cluster are considered tied. Lines between systems indicate clusters according to Wilcoxon rank-sum test at p-level $p < 0.05$. Systems with gray background indicate use of resources that fall outside the constraints provided for the shared task. NEWSTEST2018-REF denotes the human system comprised of human-produced reference translations.

As can be seen from clusters in Table 9, one system, CUNI-TRANSFORMER, appears to achieve quality better than that of the human reference, NEWSTEST2018-REF, while another, UEDIN, appears to be on par with human performance, and although both systems certainly achieve very impressive results, claims of *human parity* should be taken with a degree of caution for several reasons which we outline below.

3.7 Considerations as to Human Parity

Before making any statements about "machine translation outperforming humans" or "machine-human parity in translation", it is important to consider the following points:

- The alternate HIT structure applied in this

version of DA has not been tested thoroughly enough to be certain of high reliability. For example, as described in Section 3.3.2, forcing all translations of a given source segment to be assessed by the same human judge within the same HIT could cause individual DA ratings to become highly relative as opposed to the aim of DA ratings to be as close as possible to absolute judgments of translation quality. Furthermore, an additional bias that could cause problems for this HIT structure is one associated with a past evaluation method, relative ranking. When evaluating competing translations of the same source that are situated in close proximity within a HIT, annotators may be primed by high (or low) quality outputs resulting in overly severe (or lenient) judgments for subsequent translations of the same source segment (Bojar et al., 2011).

- While standard monolingual DA employs annotators only required to be speakers of a single language, source-based DA requires fluency in two languages and it is not known the degree to which varying levels of native language fluency in at least one language may negatively impact the reliability of DA rankings in the case of bilingual annotators.

- It is likely that the quality of reference translations can vary and this could potentially impact the reliability of human performance estimates in source-based DA. Although reference-based DA assumes high quality reference translations, in the unfortunate case of problematic references, the overall rankings are unlikely to suffer to any large degree in terms of the reliability of system rankings, since all competing systems are likely to suffer equally from any lack of quality in reference translations.

  However, in the adapted source-based version of DA, the effect of low quality reference translations is quite different. Firstly, since assessment involves comparison of MT outputs with the source, genuine participating systems will not suffer from the fact reference translations are low quality, since references are not involved in their evaluation. On the other hand, human performance estimates certainly will, as a drop in reference quality is indeed highly likely to negatively impact the placement of human performance estimates in system rankings. The reliability of comparisons with human performance with source-based DA is therefore highly dependent on high quality reference translations, as employment of a low quality set of references can only lead to *underestimates of human performance*. Considering the manual evaluation included several reports of ill-formed reference translations, conclusions of human parity and/or superiority relative to humans should be avoided.

- Since none of WMT18 systems process larger units than individual sentences and our evaluation does not include any context beyond individual segments, it is possible that the human estimate is under-rewarded for correct cross-sentential phenomena.

- The sample size employed in the source-based DA evaluation was smaller than the recommended 1,500 judgments per system.

- The way in which translations in the test sets were originally created was as follows: one half of the test data for a given language pair was translated in one language direction and the other half in the opposite direction. It is well known that the translation direction affects translation quality in training and this could also be the case for evaluation. For instance, the human reference can be scored lower for "adding" information in the case when it was actually the source sentence and the translator omitted the information when creating the translation which now serves as the source side in the test set.

- The formal education in linguistics or translatology of human assessors has not been taken into account: it is likely that whether or not human assessors have received any formal training in translation might influence their acceptance of varying levels of well-formedness in translations. For example, untrained assessors might not be as sensitive to subtle differences in verb conjugation, based on their own experience: In many real-life situations, the exact verb tense or conditional chosen in one sentence may not really impact the overall message because it can be

Test Suite	Languages	Sentences	Team
WSD-DE	en→de	3249	Rios et al. (2018)
GERMAN-LINGUISTIC	de→en	22240	Macketanz et al. (2018)
MORPHEVAL	en→{cs,de}	26500	Burlot et al. (2018)
	en→fi	16000	
	tr→en	4800	
OUT-OF-DOMAIN	en↔tr	10	Biçici (2018)
SOME-SYNTAX-PHENOMENA	en→cs	5150	Cinková and Bojar (2018)
EVALD	en→cs	2988	Bojar et al. (2018)

Table 10: Test suites employed in WMT18.

implied from the context (and thus left free to the imagination of the annotator in our sentence-based evaluation) or from general knowledge.

In sum, while we are confident that our source-based evaluation was carried out correctly, we see it only as a pilot and with conclusion limited to the very particular evaluation setting. This pilot however clearly suggests that for well-resourced language pairs, an update of WMT evaluation style will be needed to keep up with the progress in machine translation.

4 Test Suites

Arguably, both the manual and automatic evaluations carried out at WMT News Translation Task are rather opaque. We learn (for each language pair and with a known confidence) which systems perform better *on average* over the sentences sampled from the news test set.

This average performance however does not provide any insight into *which particular phenomena* are handled better or worse by the systems. It is quite possible that the overall best-performing system may be unreliable for long sentences, for named entities, for pronouns or others. Such targeted evaluations may be important for particular deployment settings and use cases, and they are definitely important for us, MT system developers, in order to focus on them in subsequent research.

To this end, WMT18 organizers ran a "call for test suites", asking researchers to design and provide sets of sentences focusing on phenomena of their interest. Table 10 lists the participating test suites and their authors. Most of the test suites were available only for a limited number of language pairs.

Each participating test-suite team provided a set of source sentences (organized into full documents, if relevant for the particular test suite). In some cases, reference translations were also made available to WMT18 organizers (but not to translation system teams).

We included the source sentences of the test suite in the source texts distributed to News Translation Task participants and collected translations of their MT systems. These, in turn, were handed over to test suite authors for evaluation. In some cases, the evaluation was fully automatic, in some cases, extended manual evaluation was carried out by the test suite team.

It is important to note that the test suite texts do not always adhere to the news domain. News Task systems which are heavily optimized towards this domain may thus underperform on such test suites. As long as this mismatch is taken into consideration, such an evaluation is valid and interesting, because it tests also the cross-domain applicability of WMT18 systems.

4.1 Test Suite Details

We now briefly describe each of the participating test suites. More details and the actual evaluation on the given test suite is available in the respective test suite paper.

4.1.1 Word Sense Disambiguation (Rios et al., 2018)

The test suite by Rios et al. (2018) presents German→English MT systems with sentences containing one of 20 German words that need to be disambiguated when translating into English, e.g. *Schlange* which could mean either a snake or a queue.

The results on that test suite clearly document that the performance in word sense disambiguation (WSD) has substantially improved over time

since 2016. While the performance in WSD generally correlates with BLEU very well, some exceptions are found, e.g. UEDIN-NMT systems from WMT16 and WMT17 or LMU-NMT performing slightly better in BLEU than in WSD. Another interesting observation is that the self-attentive architecture of Transformer seems to have a considerable advantage over RNN-based systems.

The unsupervised systems are among the worst producing, but this is in line with their low performance as estimated by BLEU.

4.1.2 Fine-Grained Evaluation for German-English (Macketanz et al., 2018)

The test suite used by Macketanz et al. (2018) is a manually designed set of 5,000 sentences covering 106 linguistic phenomena in 14 categories. The performance on this test suite is evaluated semi-automatically, with automatic checks accepting and rejecting some translations and a human annotator resolving the rest.

The results highlight the overall performance of UCAM, followed by NTT and MLLP-UPV. RWTH, JHU and UEDIN are the next group.

4.1.3 Morpheval (Burlot et al., 2018)

Burlot et al. (2018) apply the Morpheval test suite (Burlot and Yvon, 2017) and its variations to WMT18 systems translating into Czech, German, Finnish, and a smaller similar test suite also to Turkish-to-English systems. The test suite is evaluated semi-automatically and tests selected phenomena primarily reflected in morphology of Czech, German, Finnish and Turkish, resp. The tests check if translation preserves a certain contrast (e.g. the gender or number of pronouns, definiteness, verb tense or person), whether the agreement is correctly preserved under some alternation of the input (e.g. a pronoun replaced by an adjective and noun) and similarly if a particular feature is preserved across lexical variation (using a hyponym). The English-Finnish set also considers rare words: numbers and named entities in particular.

The results for English-to-Czech suggest that the Transformer model (CUNI-TRANSFORMER) may tend to produce more creative translations than the recurrent architecture (UEDIN), because it performs slightly worse in contrast preservation, most notably verb past tense, conditional, or comparative adjectives.

The English-to-German results primarily indicate that current state-of-the-art systems have no longer any real problems with internal agreement in noun phrases, coordinated verbs, preserving negation, pronoun number, strong/weak adjectives or superlatives. Phenomena like coreference, compound generation or verb future generation remain a challenge.

The English-to-Finnish evaluation again confirms easy phenomena (e.g. verb negation or preservation of numbers) and highlight language-specific hard phenomena (subordinate clause type, verb future or determiner definiteness). For this language pair, a more thorough manual validation of the test suite was also performed, indicating lower reliability for some phenomena.

Rare words (names entities) are best handled by online systems, which are probably either trained on more varied data, or include specific mechanisms to deal with this type of input, which is of lower concern for research systems.

The Turkish-English tests suggest that none of the systems handles verb particles well, with the accuracy of reflecting e.g. present vs. future subject particle in less that one third of cases. Tested verb features are handled better but apparently still considerably worse than in the other language pairs, with e.g. negation reaching only 70%.

In general, the overall performance according to human evaluation is not necessarily reflected in the performance in the Morpheval tests. A particularly interesting is the case of FACEBOOK-FAIR ⋆ (denoted "online-Z" in Burlot et al., 2018), the top English-to-German system according to manual evaluation, which performs worst in the Morpheval test on preserving morphological features under lexical variation.

4.1.4 Turkish Out-of-Domain Test (Biçici, 2018)

The set suite by Biçici (2018) consisting of only 10 sentences aimed to test the performance of English↔Turkish systems out of their news domain. Due to the small size of the test suite, it is difficult to draw any conclusions from it.

4.1.5 Czech-English Grammatical Contrasts (Cinková and Bojar, 2018)

On a set of about 3000 selected sentences (a subset of the 5150 distributed to news task participants), Cinková and Bojar (2018) examine the extent to which reference translations and MT outputs fol-

low the most prototypical pattern for certain linguistic phenomena in English-to-Czech translation. The examined MT systems include both primary News Translation Task systems, as well as three phrase-based baseline systems (Kocmi et al., 2018).

While the test suite cannot be used rank systems according to their "translation quality", it displays interesting differences among system types and the reference translation.

In essence, English control and gerund constructions can be translated as Czech finite, non-finite or subordinate clauses. The test suite focuses on cases when the particular target construction can be *expected*. According to an automatic evaluation, the reference translation follows this expected choice in about 90% of sentences of the test suite while all MT systems score considerably lower. The Moses baseline, ONLINE-G and ONLINE-A are the lowest, taking the expected route only in about 50% of cases. The top-performing system in terms of WMT18 manual evaluation, CUNI-TRANSFORMER and UEDIN perform "best" in this test suite, reaching about 70%, closely followed by the hybrid (non-primary) system Chimera (Kocmi et al., 2018) and ONLINE-B.

These results may be related e.g. to the effects of "translationese", i.e. particular constructions that appear in the target text as an artifact of the translation from a given source language. At the same time, the relation to the translation quality (see esp. Section 3.6) and the test suite results of Cinková and Bojar (2018) can be quite intricate. It is conceivable that the reference displays most of the translationese effects, CUNI-TRANSFORMER and UEDIN are able to escape this pitfall but for further systems, the scores start indicating simply a lower translation quality.

4.1.6 EVALD Discourse Evaluation (Bojar et al., 2018)

Bojar et al. (2018) present another open-ended test suite. They provide News Task systems with texts from the area of academic writing in Humanities and Arts, Social Sciences, Biological and Health Sciences, and finally Physical Sciences. After the automatic translation by WMT18 News Tasks MT systems, an automatic evaluation tool, EVALD, is used to assess the quality of the discourse.

EVALD is trained either to evaluate texts by Czech native speakers or by second-language learners. The version for Czech natives is not sufficiently discerning when applied to MT outputs, but the version for Czech learners displays measurable differences.

Since no reference is available and no manual evaluation of the machine translated texts was carried out, Bojar et al. (2018) restrict their examination to the variance of EVALD scores across subsets of the test suite. The nativeness of the original author seems to play the most important role, followed by the MT system identity and, with some gap, the genre and the domain of the text. These are promising results, confirming again that current MT systems are getting to the level of translation quality where it makes sense to compare them with tests designed for *human* writers. The quality of the source will however become the prime factor in this evaluation, only followed by the quality of the MT system.

5 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 measuring the reliability of volunteer researchers as assessors of translation quality (as opposed to crowd workers), a pilot in source-based DA evaluation and additional test suites that shed some light at the differences of individual participating MT systems and make first steps in new avenues of evaluating MT outputs using tests originally designed for humans.

In addition to highlighting the best-performing systems in each of the 14 examined translation directions, the results indicate that for some language pairs, the state of the art in machine translation is very close to the performance of human translators. This results is in line with other recent studies, e.g. Wu et al. (2016); Hassan et al. (2018), but the style of evaluation (DA for individual sentences) has to be carefully considered before making any strong claims.

Acknowledgments

This work was supported in part by funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nos. 645452 (QT21)

and 645357 (Cracker), and from the Connecting Europe Facility under agreement No. NEA/CEF/ICT/A2016/1331648 (ParaCrawl).

We would also like to thank the University of Helsinki, the University of Tartu,[18] Yandex and Microsoft for supplying test data for the news translation task.

The human evaluation campaign was very gratefully supported by contributions from Amazon, 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 would also like to give special thanks to the small group of Turkish speakers who rescued our English-Turkish human evaluation at very short notice by contributing their time voluntarily. Finally, we are grateful to the large number of anonymous Mechanical Turk workers who contributed their human intelligence to the human evaluation.

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*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv e-prints*, abs/1409.0473. Presented at ICLR 2015.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the Third Shared Task on Multimodal Machine Translation. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Chao Bei, Hao Zong, Yiming Wang, Baoyong Fan, Shiqi Li, and Conghu Yuan. 2018. An Empirical Study of Machine Translation for the Shared Task of WMT18. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Ergun Biçici. 2018. Robust parfda Statistical Machine Translation Results. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

[18]Institutional research funding IUT (20-56) of the Estonian Ministry of Education and Research.

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, Miloš Ercegovčević, Martin Popel, and Omar Zaidan. 2011. A Grain of Salt for the WMT Manual Evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 1–11, Edinburgh, Scotland. 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. 2016a. Findings of the 2016 Conference on Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

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

Ondřej Bojar, Ondřej Dušek, Tom Kocmi, Jindřich Libovický, Michal Novák, Martin Popel, Roman Sudarikov, and Dušan Variš. 2016b. CzEng 1.6: Enlarged Czech-English Parallel Corpus with Processing Tools Dockered. In *Text, Speech and Dialogue: 19th International Conference, TSD 2016, Brno,*

Czech Republic, September 12-16, 2016, Proceedings. Springer Verlag.

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

Franck Burlot, Yves Scherrer, Vinit Ravishankar, Ondřej Bojar, Stig-Arne Grönroos, Maarit Koponen, Tommi Nieminen, and François Yvon. 2018. The WMT'18 Morpheval test suites for English-Czech, English-German, English-Finnish and Turkish-English. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Franck Burlot and François Yvon. 2017. Evaluating the morphological competence of Machine Translation Systems. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. 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, Carlos Escolano, Marta R. Costa-jussà, and José A. R. Fonollosa. 2018. The TALP-UPC Machine Translation Systems for WMT18 News Shared Translation Task. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Rajen Chatterjee, Matteo Negri, Raphael Rubino, and Marco Turchi. 2018. Findings of the WMT 2018 Shared Task on Automatic Post-Editing. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Silvie Cinková and Ondřej Bojar. 2018. Testsuite on Czech–English Grammatical Contrasts. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word Translation Without Parallel Data. *CoRR*, abs/1710.04087.

Maksym Del, Andre Tättar, and Mark Fishel. 2018. Phrase-based Unsupervised Machine Translation with Compositional Phrase Embeddings. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Yongchao Deng, Shanbo Cheng, Jun Lu, Kai Song, Jingang Wang, Shenglan Wu, Liang Yao, Guchun Zhang, Haibo Zhang, Pei Zhang, Changfeng Zhu, and Boxing Chen. 2018. Alibaba's Neural Machine Translation Systems for WMT18. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. 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.

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

Markus Freitag, Minwei Feng, Matthias Huck, Stephan Peitz, and Hermann Ney. 2013. Reverse Word Order Models. In *Proceedings of the XIV Machine Translation Summit*, pages 159–166, Nice, France.

Markus Freitag, Matthias Huck, and Hermann Ney. 2014. Jane: Open Source Machine Translation System Combination. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 29–32, Gothenburg, Sweden. Association for Computational Linguistics.

Adrià de Gispert, Bill Byrne, Eva Hasler, and Felix Stahlberg. 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, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 362–368.

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.

Miguel Graça, Yunsu Kim, Julian Schamper, Jiahui Geng, and Hermann Ney. 2018. The RWTH Aachen University English-German and German-English Unsupervised Neural Machine Translation Systems for WMT 2018. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Stig-Arne Grönroos, Sami Virpioja, and Mikko Kurimo. 2018. Cognate-aware morphological segmentation for multilingual neural translation. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Jeremy Gwinnup, Tim Anderson, Grant Erdmann, and Katherine Young. 2018. The AFRL WMT18 Systems: Ensembling, Continuation and Combination. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Jeremy Gwinnup, Timothy Anderson, Grant Erdmann, Katherine Young, Michaeel Kazi, Elizabeth Salesky, Brian Thompson, and Jonathan Taylor. 2017. The AFRL-MITLL WMT17 Systems: Old, New, Borrowed, BLEU. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Barry Haddow, Nikolay Bogoychev, Denis Emelin, Ulrich Germann, Roman Grundkiewicz, Kenneth Heafield, Antonio Valerio Miceli Barone, and Rico Sennrich. 2018. The University of Edinburgh's Submissions to the WMT18 News Translation Task. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

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.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A Toolkit for Neural Machine Translation. *arXiv preprint arXiv:1712.05690*.

Bojie Hu, Ambyer Han, and Shen Huang. 2018. TencentFmRD Neural Machine Translation for WMT18. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Matthias Huck, Fabienne Braune, and Alexander Fraser. 2017a. LMU Munich's Neural Machine Translation Systems for News Articles and Health Information Texts. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Matthias Huck, Simon Riess, and Alexander Fraser. 2017b. Target-side Word Segmentation Strategies for Neural Machine Translation. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Matthias Huck, Dario Stojanovski, Viktor Hangya, and Alexander Fraser. 2018. LMU Munich's Neural Machine Translation Systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Arvi Hurskainen and Jörg Tiedemann. 2017. Rule-based Machine translation from English to Finnish. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Javier Iranzo-Sánchez, Pau Baquero-Arnal, Gonçal V. Garcés Díaz-Munío, Adrià Martínez-Villaronga, Jorge Civera, and Alfons Juan. 2018. The MLLP-UPV German-English Machine Translation System for WMT18. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2018. 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, 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.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Tom Kocmi, Roman Sudarikov, and Ondřej Bojar. 2018. CUNI Submissions in WMT18. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Philipp Koehn, Kevin Duh, and Brian Thompson. 2018a. 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, 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, Huda Khayrallah, Kenneth Heafield, and Mikel L. Forcada. 2018b. Findings of the WMT 2018 Shared Task on Parallel Corpus Filtering. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. 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.

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

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-Based & Neural Unsupervised Machine Translation. *arXiv preprint arXiv:1804.07755*.

Lemao Liu, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. 2016. Agreement on Target-bidirectional Neural Machine Translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 411–416, San Diego, CA, USA. Association for Computational Linguistics.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, Brussels, Belgium. 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.

Sainik Kumar Mahata, Dipankar Das, and Sivaji Bandyopadhyay. 2018. JUCBNMT at WMT2018 News Translation Task: Character Based Neural Machine Translation of Finnish to English. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Benjamin Marie, Rui Wang, Atsushi Fujita, Masao Utiyama, and Eiichiro Sumita. 2018. NICT's Neural and Statistical Machine Translation Systems for the WMT18 News Translation Task. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Alexander Molchanov. 2018. PROMT Systems for WMT 2018 Shared Translation Task. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2018. NTT's Neural Machine Translation Systems for WMT 2018. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Mariana Neves, Antonio Jimeno Yepes, Aurélie Névéol, Cristian Grozea, Amy Siu, Madeleine Kittner, and Karin Verspoor. 2018. Findings of the WMT 2018 Biomedical Translation Shared Task: Evaluation on Medline test sets. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Ngoc-Quan Pham, Jan Niehues, and Alexander Waibel. 2018. The Karlsruhe Institute of Technology Systems for the News Translation Task in WMT 2018. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Mārcis Pinnis, Matīss Rikters, and Rihards Krišlauks. 2018. Tilde's Machine Translation Systems for WMT 2018. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. 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 and Ondřej Bojar. 2018. Training Tips for the Transformer Model. *The Prague Bulletin of Mathematical Linguistics*, 110(1):43–70.

Alessandro Raganato, Yves Scherrer, Tommi Nieminen, Arvi Hurskainen, and Jörg Tiedemann. 2018. The University of Helsinki submissions to the WMT18 news task. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Annette Rios, Mathias Müller, and Rico Sennrich. 2018. The Word Sense Disambiguation Test Suite at WMT18. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Julian Schamper, Jan Rosendahl, Parnia Bahar, Yunsu Kim, Arne Nix, and Hermann Ney. 2018. The RWTH Aachen University Supervised Machine Translation Systems for WMT 2018. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017. Nematus: a Toolkit for Neural Machine Translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

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

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón Astudillo, and André F. T. Martins. 2018. Findings of the WMT 2018 Shared Task on Quality Estimation. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Felix Stahlberg, Adrià de Gispert, and Bill Byrne. 2018. The University of Cambridge's Machine Translation Systems for WMT18. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Dario Stojanovski, Viktor Hangya, Matthias Huck, and Alexander Fraser. 2018. The LMU Munich Unsupervised Machine Translation Systems. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. 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.

Sander Tars and Mark Fishel. 2018. Multi-Domain Neural Machine Translation. In *Proceedings of EAMT*, pages 259–268.

Jörg Tiedemann, Fabienne Cap, Jenna Kanerva, Filip Ginter, Sara Stymne, Robert Östling, and Marion Weller-Di Marco. 2016. Phrase-Based SMT for Finnish with More Data, Better Models and Alternative Alignment and Translation Tools. In *Proceedings of the First Conference on Machine Translation*, pages 391–398, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2Tensor for Neural Machine Translation. *CoRR*, abs/1803.07416.

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.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, and Mikko Kurimo. 2013. Morfessor 2.0: Python Implementation and Extensions for Morfessor Baseline. Technical report.

Mingxuan Wang, Li Gong, Wenhuan Zhu, Jun Xie, and Chao Bian. 2018a. Tencent Neural Machine Translation Systems for WMT18. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Qiang Wang, Bei Li, Jiqiang Liu, Bojian Jiang, Zheyang Zhang, Yinqiao Li, Ye Lin, Tong Xiao, and Jingbo Zhu. 2018b. The NiuTrans Machine Translation System for WMT18. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Weijia Xu and Marine Carpuat. 2018. The University of Maryland's Chinese-English Neural Machine Translation Systems at WMT18. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium. Association for Computational Linguistics.

A Differences in Human Scores

Tables 11–24 show differences in average standardized human scores for all pairs of competing systems for each language pair. The numbers in each of the tables' cells indicate the difference in average standardized human scores for the system in that column and the system in that row.

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

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

Table 25 shows the differences in average standardized human scores for Czech→English systems, based on source-based DA.

	NiuTrans	online-B	UCAM	Unisound-A	Tencent-ensemble	Unisound-B	Li-Muze	NICT	UMD	online-Y	uedin	online-A	online-G	online-F
NiuTrans	-	0.03	0.03	0.03	0.04⋆	0.05⋆	0.05⋆	0.05⋆	0.06†	0.15‡	0.16‡	0.20‡	0.47‡	0.52‡
online-B	-0.03	-	0.00	0.00	0.01	0.02	0.02	0.02	0.03	0.12‡	0.13‡	0.17‡	0.44‡	0.49‡
UCAM	-0.03	0.00	-	0.00	0.01	0.02	0.02	0.02	0.03	0.11‡	0.13‡	0.17‡	0.44‡	0.49‡
Unisound-A	-0.03	0.00	0.00	-	0.01	0.01	0.02	0.02	0.03	0.11‡	0.12‡	0.17‡	0.43‡	0.48‡
Tencent-ensemble	-0.04	-0.01	-0.01	-0.01	-	0.01	0.01	0.01	0.02	0.10‡	0.12‡	0.16‡	0.43‡	0.48‡
Unisound-B	-0.05	-0.02	-0.02	-0.01	-0.01	-	0.00	0.01	0.02	0.10‡	0.11‡	0.16‡	0.42‡	0.47‡
Li-Muze	-0.05	-0.02	-0.02	-0.02	-0.01	0.00	-	0.00	0.01	0.10‡	0.11‡	0.15‡	0.42‡	0.47‡
NICT	-0.05	-0.02	-0.02	-0.02	-0.01	-0.01	0.00	-	0.01	0.09‡	0.11‡	0.15‡	0.42‡	0.47‡
UMD	-0.06	-0.03	-0.03	-0.03	-0.02	-0.02	-0.01	-0.01	-	0.08†	0.10‡	0.14‡	0.40‡	0.45‡
online-Y	-0.15	-0.12	-0.11	-0.11	-0.10	-0.10	-0.10	-0.09	-0.08	-	0.01	0.06⋆	0.32‡	0.37‡
uedin	-0.16	-0.13	-0.13	-0.12	-0.12	-0.11	-0.11	-0.11	-0.10	-0.01	-	0.04⋆	0.31‡	0.36‡
online-A	-0.20	-0.17	-0.17	-0.17	-0.16	-0.16	-0.15	-0.15	-0.14	-0.06	-0.04	-	0.27‡	0.32‡
online-G	-0.47	-0.44	-0.44	-0.43	-0.43	-0.42	-0.42	-0.42	-0.40	-0.32	-0.31	-0.27	-	0.05⋆
online-F	-0.52	-0.49	-0.49	-0.48	-0.48	-0.47	-0.47	-0.47	-0.45	-0.37	-0.36	-0.32	-0.05	-
score	0.14	0.11	0.11	0.11	0.10	0.09	0.09	0.09	0.08	-0.01	-0.02	-0.06	-0.33	-0.38
rank	1–9	1–9	1–9	1–9	1–9	1–9	1–9	1–9	1–9	10–11	10–11	12	13	14

Table 11: Head to head comparison for Chinese→English systems.

	TENCENT-ENSEMBLE	UNISOUND	GTCOM-PRIMARY	ALIBABA-ENS-RERANK	ALIBABA-GENERAL-A	ONLINE-B	ALIBABA-GENERAL-B	UMD	NICT	ONLINE-Y	ONLINE-A	UEDIN	ONLINE-F	ONLINE-G
TENCENT-ENSEMBLE	-	0.01	0.02	0.03	0.05	0.05⋆	0.05	0.13‡	0.14‡	0.15‡	0.18‡	0.42‡	0.64‡	0.65‡
UNISOUND	-0.01	-	0.01	0.02	0.03	0.04	0.04	0.11‡	0.12‡	0.14‡	0.17‡	0.41‡	0.62‡	0.64‡
GTCOM-PRIMARY	-0.02	-0.01	-	0.01	0.03	0.03⋆	0.03	0.11‡	0.12‡	0.13‡	0.16‡	0.40‡	0.62‡	0.63‡
ALIBABA-ENS-RERANK	-0.03	-0.02	-0.01	-	0.01	0.02	0.02	0.09‡	0.10‡	0.12‡	0.15‡	0.39‡	0.60‡	0.62‡
ALIBABA-GENERAL-A	-0.05	-0.03	-0.03	-0.01	-	0.01	0.01	0.08‡	0.09†	0.10‡	0.14‡	0.38‡	0.59‡	0.61‡
ONLINE-B	-0.05	-0.04	-0.03	-0.02	-0.01	-	0.00	0.07†	0.08⋆	0.10‡	0.13‡	0.37‡	0.58‡	0.60‡
ALIBABA-GENERAL-B	-0.05	-0.04	-0.03	-0.02	-0.01	0.00	-	0.07†	0.08†	0.10‡	0.13‡	0.37‡	0.58‡	0.60‡
UMD	-0.13	-0.11	-0.11	-0.09	-0.08	-0.07	-0.07	-	0.01	0.03	0.06†	0.30‡	0.51‡	0.53‡
NICT	-0.14	-0.12	-0.12	-0.10	-0.09	-0.08	-0.08	-0.01	-	0.01	0.04†	0.28‡	0.50‡	0.52‡
ONLINE-Y	-0.15	-0.14	-0.13	-0.12	-0.10	-0.10	-0.10	-0.03	-0.01	-	0.03	0.27‡	0.49‡	0.50‡
ONLINE-A	-0.18	-0.17	-0.16	-0.15	-0.14	-0.13	-0.13	-0.06	-0.04	-0.03	-	0.24‡	0.46‡	0.47‡
UEDIN	-0.42	-0.41	-0.40	-0.39	-0.38	-0.37	-0.37	-0.30	-0.28	-0.27	-0.24	-	0.22‡	0.23‡
ONLINE-F	-0.64	-0.62	-0.62	-0.60	-0.59	-0.58	-0.58	-0.51	-0.50	-0.49	-0.46	-0.22	-	0.02
ONLINE-G	-0.65	-0.64	-0.63	-0.62	-0.61	-0.60	-0.60	-0.53	-0.52	-0.50	-0.47	-0.23	-0.02	-
score	0.22	0.21	0.20	0.18	0.17	0.17	0.17	0.09	0.08	0.07	0.04	-0.20	-0.42	-0.43
rank	1–7	1–7	1–7	1–7	1–7	1–7	1–7	8–11	8–11	8–11	8–11	12	13–14	13–14

Table 12: Head to head comparison for English→Chinese systems.

	CUNI-TRANSFORMER	UEDIN	ONLINE-B	ONLINE-A	ONLINE-G
CUNI-TRANSFORMER	-	0.13‡	0.18‡	0.32‡	0.48‡
UEDIN	-0.13	-	0.05⋆	0.19‡	0.35‡
ONLINE-B	-0.18	-0.05	-	0.14‡	0.30‡
ONLINE-A	-0.32	-0.19	-0.14	-	0.16‡
ONLINE-G	-0.48	-0.35	-0.30	-0.16	-
score	0.30	0.17	0.12	-0.02	-0.18
rank	1	2	3	4	5

Table 13: Head to head comparison for Czech→English systems.

	CUNI-TRANSFORMER	UEDIN	ONLINE-B	ONLINE-A	ONLINE-G
CUNI-TRANSFORMER	-	0.21‡	0.49‡	0.71‡	0.84‡
UEDIN	-0.21	-	0.28‡	0.50‡	0.63‡
ONLINE-B	-0.49	-0.28	-	0.22‡	0.35‡
ONLINE-A	-0.71	-0.50	-0.22	-	0.13‡
ONLINE-G	-0.84	-0.63	-0.35	-0.13	-
score	0.59	0.38	0.10	-0.12	-0.25
rank	1	2	3	4	5

Table 14: Head to head comparison for English→Czech systems.

	RWTH	UCAM	NTT	ONLINE-B	MLLP-UPV	JHU	UBIQUS-NMT	ONLINE-Y	ONLINE-A	UEDIN	LMU-NMT	NJUNMT-PRIVATE	ONLINE-G	ONLINE-F	RWTH-UNSUPER	LMU-UNSUP
RWTH	-	0.02	0.05⋆	0.07‡	0.09‡	0.10‡	0.10‡	0.10‡	0.15‡	0.15‡	0.25‡	0.26‡	0.49‡	0.71‡	1.17‡	1.25‡
UCAM	-0.02	-	0.04	0.05⋆	0.07‡	0.08†	0.08‡	0.08‡	0.13‡	0.13‡	0.23‡	0.25‡	0.47‡	0.69‡	1.15‡	1.23‡
NTT	-0.05	-0.04	-	0.01	0.04⋆	0.04⋆	0.04†	0.05†	0.09‡	0.10‡	0.20‡	0.21‡	0.43‡	0.66‡	1.11‡	1.19‡
ONLINE-B	-0.07	-0.05	-0.01	-	0.03	0.03	0.03⋆	0.04	0.08‡	0.09‡	0.18‡	0.20‡	0.42‡	0.64‡	1.10‡	1.18‡
MLLP-UPV	-0.09	-0.07	-0.04	-0.03	-	0.00	0.01	0.01	0.05†	0.06⋆	0.16‡	0.17‡	0.40‡	0.62‡	1.07‡	1.16‡
JHU	-0.10	-0.08	-0.04	-0.03	0.00	-	0.00	0.01	0.05†	0.06†	0.15‡	0.17‡	0.39‡	0.61‡	1.07‡	1.15‡
UBIQUS-NMT	-0.10	-0.08	-0.04	-0.03	-0.01	0.00	-	0.01	0.05⋆	0.05	0.15‡	0.17‡	0.39‡	0.61‡	1.07‡	1.15‡
ONLINE-Y	-0.10	-0.08	-0.05	-0.04	-0.01	-0.01	-0.01	-	0.04⋆	0.05⋆	0.15‡	0.16‡	0.38‡	0.61‡	1.06‡	1.15‡
ONLINE-A	-0.15	-0.13	-0.09	-0.08	-0.05	-0.05	-0.05	-0.04	-	0.01	0.11‡	0.12‡	0.34‡	0.56‡	1.02‡	1.10‡
UEDIN	-0.15	-0.13	-0.10	-0.09	-0.06	-0.06	-0.05	-0.05	-0.01	-	0.10‡	0.11‡	0.34‡	0.56‡	1.01‡	1.10‡
LMU-NMT	-0.25	-0.23	-0.20	-0.18	-0.16	-0.15	-0.15	-0.15	-0.11	-0.10	-	0.01	0.24‡	0.46‡	0.91‡	1.00‡
NJUNMT-PRIVATE	-0.26	-0.25	-0.21	-0.20	-0.17	-0.17	-0.17	-0.16	-0.12	-0.11	-0.01	-	0.22‡	0.45‡	0.90‡	0.98‡
ONLINE-G	-0.49	-0.47	-0.43	-0.42	-0.40	-0.39	-0.39	-0.38	-0.34	-0.34	-0.24	-0.22	-	0.22‡	0.68‡	0.76‡
ONLINE-F	-0.71	-0.69	-0.66	-0.64	-0.62	-0.61	-0.61	-0.61	-0.56	-0.56	-0.46	-0.45	-0.22	-	0.46‡	0.54‡
RWTH-UNSUPER	-1.17	-1.15	-1.11	-1.10	-1.07	-1.07	-1.07	-1.06	-1.02	-1.01	-0.91	-0.90	-0.68	-0.46	-	0.08‡
LMU-UNSUP	-1.25	-1.23	-1.19	-1.18	-1.16	-1.15	-1.15	-1.15	-1.10	-1.10	-1.00	-0.98	-0.76	-0.54	-0.08	-
score	0.41	0.40	0.36	0.35	0.32	0.32	0.32	0.31	0.27	0.26	0.16	0.15	-0.07	-0.30	-0.75	-0.83
rank	1–8	1–8	1–8	1–8	1–8	1–8	1–8	1–8	9–10	9–10	11–12	11–12	13	14	15	16

Table 15: Head to head comparison for German→English systems.

	FACEBOOK-FAIR ⋆	ONLINE-B	MICROSOFT-MARIAN	MMT-PRODUCTION	UCAM	NTT	KIT	ONLINE-Y	JHU	UEDIN	LMU-NMT	ONLINE-A	ONLINE-F	ONLINE-G	RWTH-UNSUPER	LMU-UNSUP
FACEBOOK-FAIR ⋆	-	0.09†	0.10‡	0.11‡	0.12‡	0.16‡	0.20‡	0.26‡	0.28‡	0.30‡	0.44‡	0.59‡	1.04‡	1.07‡	1.62‡	1.77‡
ONLINE-B	-0.09	-	0.01	0.02	0.02	0.07†	0.11‡	0.16‡	0.18‡	0.21‡	0.35‡	0.50‡	0.95‡	0.98‡	1.53‡	1.68‡
MICROSOFT-MARIAN	-0.10	-0.01	-	0.01	0.01	0.06⋆	0.10†	0.16‡	0.17‡	0.20‡	0.34‡	0.49‡	0.94‡	0.97‡	1.52‡	1.67‡
MMT-PRODUCTION	-0.11	-0.02	-0.01	-	0.00	0.05	0.09⋆	0.14‡	0.16‡	0.19‡	0.33‡	0.48‡	0.92‡	0.95‡	1.51‡	1.66‡
UCAM	-0.12	-0.02	-0.01	0.00	-	0.05	0.08⋆	0.14‡	0.16‡	0.19‡	0.32‡	0.48‡	0.92‡	0.95‡	1.50‡	1.66‡
NTT	-0.16	-0.07	-0.06	-0.05	-0.05	-	0.04	0.10‡	0.11†	0.14‡	0.28‡	0.43‡	0.88‡	0.91‡	1.46‡	1.61‡
KIT	-0.20	-0.11	-0.10	-0.09	-0.08	-0.04	-	0.06†	0.08⋆	0.10†	0.24‡	0.39‡	0.84‡	0.87‡	1.42‡	1.58‡
ONLINE-Y	-0.26	-0.16	-0.16	-0.14	-0.14	-0.10	-0.06	-	0.02	0.04	0.18‡	0.34‡	0.78‡	0.81‡	1.36‡	1.52‡
JHU	-0.28	-0.18	-0.17	-0.16	-0.16	-0.11	-0.08	-0.02	-	0.03	0.16‡	0.32‡	0.76‡	0.79‡	1.34‡	1.50‡
UEDIN	-0.30	-0.21	-0.20	-0.19	-0.19	-0.14	-0.10	-0.04	-0.03	-	0.14‡	0.29‡	0.74‡	0.77‡	1.32‡	1.47‡
LMU-NMT	-0.44	-0.35	-0.34	-0.33	-0.32	-0.28	-0.24	-0.18	-0.16	-0.14	-	0.15‡	0.60‡	0.63‡	1.18‡	1.33‡
ONLINE-A	-0.59	-0.50	-0.49	-0.48	-0.48	-0.43	-0.39	-0.34	-0.32	-0.29	-0.15	-	0.44‡	0.48‡	1.03‡	1.18‡
ONLINE-F	-1.04	-0.95	-0.94	-0.92	-0.92	-0.88	-0.84	-0.78	-0.76	-0.74	-0.60	-0.44	-	0.03	0.58‡	0.74‡
ONLINE-G	-1.07	-0.98	-0.97	-0.95	-0.95	-0.91	-0.87	-0.81	-0.79	-0.77	-0.63	-0.48	-0.03	-	0.55‡	0.71‡
RWTH-UNSUPER	-1.62	-1.53	-1.52	-1.51	-1.50	-1.46	-1.42	-1.36	-1.34	-1.32	-1.18	-1.03	-0.58	-0.55	-	0.16‡
LMU-UNSUP	-1.77	-1.68	-1.67	-1.66	-1.66	-1.61	-1.58	-1.52	-1.50	-1.47	-1.33	-1.18	-0.74	-0.71	-0.16	-
score	0.65	0.56	0.55	0.54	0.54	0.49	0.45	0.40	0.38	0.35	0.21	0.06	-0.39	-0.42	-0.97	-1.12
rank	1	2–7	2–7	2–7	2–7	2–7	2–7	8–10	8–10	8–10	11	12	13–14	13–14	15	16

Table 16: Head to head comparison for English→German systems.

	TILDE-NC-NMT	NICT	TILDE-C-NMT	TILDE-C-NMT-2BT	UEDIN	TILDE-C-NMT-COMB	ONLINE-B	HY-NMT	TALP-UPC	ONLINE-A	CUNI-KOCMI	NEUROTOLGE.EE	ONLINE-G	UNSUPTARTU
TILDE-NC-NMT	-	0.09‡	0.11‡	0.14‡	0.14‡	0.16‡	0.21‡	0.22‡	0.22‡	0.26‡	0.32‡	0.44‡	0.67‡	1.28‡
NICT	-0.09	-	0.02	0.05⋆	0.05	0.07†	0.12‡	0.13‡	0.13‡	0.18‡	0.23‡	0.36‡	0.58‡	1.19‡
TILDE-C-NMT	-0.11	-0.02	-	0.03	0.03	0.04	0.10‡	0.11†	0.11‡	0.15‡	0.21‡	0.33‡	0.56‡	1.17‡
TILDE-C-NMT-2BT	-0.14	-0.05	-0.03	-	0.00	0.02	0.07†	0.08†	0.08†	0.12‡	0.18‡	0.30‡	0.53‡	1.14‡
UEDIN	-0.14	-0.05	-0.03	0.00	-	0.02	0.07†	0.08†	0.08†	0.12‡	0.18‡	0.30‡	0.53‡	1.14‡
TILDE-C-NMT-COMB	-0.16	-0.07	-0.04	-0.02	-0.02	-	0.05⋆	0.06	0.06⋆	0.11‡	0.16‡	0.29‡	0.51‡	1.12‡
ONLINE-B	-0.21	-0.12	-0.10	-0.07	-0.07	-0.05	-	0.01	0.01	0.05⋆	0.11†	0.23‡	0.46‡	1.07‡
HY-NMT	-0.22	-0.13	-0.11	-0.08	-0.08	-0.06	-0.01	-	0.00	0.04⋆	0.10†	0.22‡	0.45‡	1.06‡
TALP-UPC	-0.22	-0.13	-0.11	-0.08	-0.08	-0.06	-0.01	0.00	-	0.04⋆	0.10†	0.22‡	0.45‡	1.06‡
ONLINE-A	-0.26	-0.18	-0.15	-0.12	-0.12	-0.11	-0.05	-0.04	-0.04	-	0.06	0.18‡	0.40‡	1.01‡
CUNI-KOCMI	-0.32	-0.23	-0.21	-0.18	-0.18	-0.16	-0.11	-0.10	-0.10	-0.06	-	0.12‡	0.35‡	0.96‡
NEUROTOLGE.EE	-0.44	-0.36	-0.33	-0.30	-0.30	-0.29	-0.23	-0.22	-0.22	-0.18	-0.12	-	0.22‡	0.83‡
ONLINE-G	-0.67	-0.58	-0.56	-0.53	-0.53	-0.51	-0.46	-0.45	-0.45	-0.40	-0.35	-0.22	-	0.61‡
UNSUPTARTU	-1.28	-1.19	-1.17	-1.14	-1.14	-1.12	-1.07	-1.06	-1.06	-1.01	-0.96	-0.83	-0.61	-
score	0.33	0.24	0.21	0.19	0.19	0.17	0.12	0.11	0.11	0.06	0.01	-0.12	-0.34	-0.95
rank	1	2–9	2–9	2–9	2–9	2–9	2–9	2–9	2–9	10–11	10–11	12	13	14

Table 17: Head to head comparison for Estonian→English systems.

	TILDE-NC-NMT	NICT	TILDE-C-NMT	TILDE-C-NMT-2BT	AALTO	HY-NMT	UEDIN	CUNI-KOCMI	TALP-UPC	ONLINE-B	NEUROTOLGE.EE	ONLINE-A	ONLINE-G	PARFDA
TILDE-NC-NMT	-	0.10⋆	0.12†	0.13‡	0.21‡	0.22‡	0.25‡	0.33‡	0.37‡	0.45‡	0.68‡	0.74‡	0.95‡	1.07‡
NICT	-0.10	-	0.03	0.03	0.11†	0.12†	0.16‡	0.24‡	0.27‡	0.36‡	0.58‡	0.65‡	0.86‡	0.97‡
TILDE-C-NMT	-0.12	-0.03	-	0.01	0.09⋆	0.10⋆	0.13†	0.21‡	0.25‡	0.33‡	0.56‡	0.62‡	0.83‡	0.95‡
TILDE-C-NMT-2BT	-0.13	-0.03	-0.01	-	0.08⋆	0.09⋆	0.12†	0.20‡	0.24‡	0.32‡	0.55‡	0.61‡	0.82‡	0.94‡
AALTO	-0.21	-0.11	-0.09	-0.08	-	0.01	0.05	0.12†	0.16‡	0.24‡	0.47‡	0.54‡	0.75‡	0.86‡
HY-NMT	-0.22	-0.12	-0.10	-0.09	-0.01	-	0.03	0.11†	0.15‡	0.23‡	0.46‡	0.52‡	0.73‡	0.85‡
UEDIN	-0.25	-0.16	-0.13	-0.12	-0.05	-0.03	-	0.08⋆	0.11†	0.20‡	0.43‡	0.49‡	0.70‡	0.81‡
CUNI-KOCMI	-0.33	-0.24	-0.21	-0.20	-0.12	-0.11	-0.08	-	0.04	0.12†	0.35‡	0.41‡	0.62‡	0.74‡
TALP-UPC	-0.37	-0.27	-0.25	-0.24	-0.16	-0.15	-0.11	-0.04	-	0.08⋆	0.31‡	0.38‡	0.59‡	0.70‡
ONLINE-B	-0.45	-0.36	-0.33	-0.32	-0.24	-0.23	-0.20	-0.12	-0.08	-	0.23‡	0.29‡	0.50‡	0.62‡
NEUROTOLGE.EE	-0.68	-0.58	-0.56	-0.55	-0.47	-0.46	-0.43	-0.35	-0.31	-0.23	-	0.06⋆	0.27‡	0.39‡
ONLINE-A	-0.74	-0.65	-0.62	-0.61	-0.54	-0.52	-0.49	-0.41	-0.38	-0.29	-0.06	-	0.21‡	0.32‡
ONLINE-G	-0.95	-0.86	-0.83	-0.82	-0.75	-0.73	-0.70	-0.62	-0.59	-0.50	-0.27	-0.21	-	0.11‡
PARFDA	-1.07	-0.97	-0.95	-0.94	-0.86	-0.85	-0.81	-0.74	-0.70	-0.62	-0.39	-0.32	-0.11	-
score	0.55	0.45	0.43	0.42	0.34	0.33	0.29	0.22	0.18	0.10	-0.13	-0.20	-0.41	-0.52
rank	1	2–4	2–4	2–4	5–7	5–7	5–7	8–9	8–9	10	11	12	13	14

Table 18: Head to head comparison for English→Estonian systems.

	NICT	HY-NMT	UEDIN	CUNI-KOCMI	ONLINE-B	TALP-UPC	ONLINE-A	ONLINE-G	JUCBNMT
NICT	-	0.02	0.05	0.07†	0.07†	0.11‡	0.11‡	0.29‡	0.56‡
HY-NMT	-0.02	-	0.03	0.05⋆	0.05⋆	0.08†	0.08†	0.26‡	0.53‡
UEDIN	-0.05	-0.03	-	0.02	0.02	0.06†	0.06†	0.24‡	0.51‡
CUNI-KOCMI	-0.07	-0.05	-0.02	-	0.00	0.04	0.04	0.22‡	0.49‡
ONLINE-B	-0.07	-0.05	-0.02	0.00	-	0.03	0.03	0.21‡	0.48‡
TALP-UPC	-0.11	-0.08	-0.06	-0.04	-0.03	-	0.00	0.18‡	0.45‡
ONLINE-A	-0.11	-0.08	-0.06	-0.04	-0.03	0.00	-	0.18‡	0.45‡
ONLINE-G	-0.29	-0.26	-0.24	-0.22	-0.21	-0.18	-0.18	-	0.27‡
JUCBNMT	-0.56	-0.53	-0.51	-0.49	-0.48	-0.45	-0.45	-0.27	-
score	0.15	0.13	0.10	0.08	0.08	0.05	0.04	-0.13	-0.40
rank	1–7	1–7	1–7	1–7	1–7	1–7	1–7	8	9

Table 19: Head to head comparison for Finnish→English systems.

	NICT	HY-NMT	UEDIN	AALTO	HY-NMT-2STEP	TALP-UPC	CUNI-KOCMI	ONLINE-B	ONLINE-A	ONLINE-G	HY-SMT	HY-AH
NICT	-	0.05	0.20‡	0.25‡	0.26‡	0.28‡	0.34‡	0.34‡	0.73‡	0.75‡	0.86‡	0.89‡
HY-NMT	-0.05	-	0.14†	0.19‡	0.21‡	0.23‡	0.28‡	0.28‡	0.68‡	0.70‡	0.80‡	0.83‡
UEDIN	-0.20	-0.14	-	0.05	0.07	0.09⋆	0.14†	0.14‡	0.54‡	0.56‡	0.66‡	0.69‡
AALTO	-0.25	-0.19	-0.05	-	0.01	0.03	0.09⋆	0.09⋆	0.48‡	0.50‡	0.61‡	0.64‡
HY-NMT-2STEP	-0.26	-0.21	-0.07	-0.01	-	0.02	0.07	0.07⋆	0.47‡	0.49‡	0.59‡	0.63‡
TALP-UPC	-0.28	-0.23	-0.09	-0.03	-0.02	-	0.05	0.05	0.45‡	0.47‡	0.57‡	0.61‡
CUNI-KOCMI	-0.34	-0.28	-0.14	-0.09	-0.07	-0.05	-	0.00	0.40‡	0.42‡	0.52‡	0.55‡
ONLINE-B	-0.34	-0.28	-0.14	-0.09	-0.07	-0.05	0.00	-	0.39‡	0.42‡	0.52‡	0.55‡
ONLINE-A	-0.73	-0.68	-0.54	-0.48	-0.47	-0.45	-0.40	-0.39	-	0.02	0.12†	0.16‡
ONLINE-G	-0.75	-0.70	-0.56	-0.50	-0.49	-0.47	-0.42	-0.42	-0.02	-	0.10†	0.14‡
HY-SMT	-0.86	-0.80	-0.66	-0.61	-0.59	-0.57	-0.52	-0.52	-0.12	-0.10	-	0.03
HY-AH	-0.89	-0.83	-0.69	-0.64	-0.63	-0.61	-0.55	-0.55	-0.16	-0.14	-0.03	-
score	0.52	0.47	0.32	0.27	0.26	0.24	0.18	0.18	-0.21	-0.23	-0.33	-0.37
rank	1–2	1–2	3–8	3–8	3–8	3–8	3–8	3–8	9–10	9–10	11–12	11–12

Table 20: Head to head comparison for English→Finnish systems.

	ALIBABA	ONLINE-B	ONLINE-G	UEDIN	ONLINE-A	AFRL-SYSCOMB	JHU	ONLINE-F
ALIBABA	-	0.02	0.04	0.10‡	0.18‡	0.23‡	0.24‡	0.61‡
ONLINE-B	-0.02	-	0.02	0.08⋆	0.16‡	0.21‡	0.22‡	0.59‡
ONLINE-G	-0.04	-0.02	-	0.06⋆	0.14‡	0.18‡	0.20‡	0.57‡
UEDIN	-0.10	-0.08	-0.06	-	0.08†	0.12‡	0.14‡	0.51‡
ONLINE-A	-0.18	-0.16	-0.14	-0.08	-	0.05⋆	0.06⋆	0.43‡
AFRL-SYSCOMB	-0.23	-0.21	-0.18	-0.12	-0.05	-	0.01	0.38‡
JHU	-0.24	-0.22	-0.20	-0.14	-0.06	-0.01	-	0.37‡
ONLINE-F	-0.61	-0.59	-0.57	-0.51	-0.43	-0.38	-0.37	-
score	0.21	0.19	0.17	0.11	0.03	-0.01	-0.03	-0.40
rank	1–3	1–3	1–3	4	5	6–7	6–7	8

Table 21: Head to head comparison for Russian→English systems.

	ALIBABA-ENS	ONLINE-G	ONLINE-B	UEDIN	PROMT-HYB-MARIAN	PROMT-HYB-OPENNMT	ONLINE-A	PROMT-RULE-BASED	ONLINE-F
ALIBABA-ENS	-	0.03	0.19‡	0.21‡	0.24‡	0.29‡	0.36‡	0.43‡	0.93‡
ONLINE-G	-0.03	-	0.16‡	0.18‡	0.21‡	0.26‡	0.33‡	0.40‡	0.90‡
ONLINE-B	-0.19	-0.16	-	0.01	0.04⋆	0.09‡	0.16‡	0.23‡	0.74‡
UEDIN	-0.21	-0.18	-0.01	-	0.03	0.08†	0.15‡	0.22‡	0.72‡
PROMT-HYB-MARIAN	-0.24	-0.21	-0.04	-0.03	-	0.05⋆	0.12‡	0.19‡	0.69‡
PROMT-HYB-OPENNMT	-0.29	-0.26	-0.09	-0.08	-0.05	-	0.07†	0.14‡	0.65‡
ONLINE-A	-0.36	-0.33	-0.16	-0.15	-0.12	-0.07	-	0.07†	0.58‡
PROMT-RULE-BASED	-0.43	-0.40	-0.23	-0.22	-0.19	-0.14	-0.07	-	0.50‡
ONLINE-F	-0.93	-0.90	-0.74	-0.72	-0.69	-0.65	-0.58	-0.50	-
score	0.35	0.32	0.16	0.14	0.12	0.07	-0.00	-0.07	-0.58
rank	1–2	1–2	3–5	3–5	3–5	6	7	8	9

Table 22: Head to head comparison for English→Russian systems.

	ONLINE-G	ONLINE-A	ALIBABA-ENS	ONLINE-B	UEDIN	NICT
ONLINE-G	-	0.02	0.06⋆	0.06†	0.11‡	0.13‡
ONLINE-A	-0.02	-	0.04	0.04	0.08⋆	0.10‡
ALIBABA-ENS	-0.06	-0.04	-	0.01	0.05	0.07†
ONLINE-B	-0.06	-0.04	-0.01	-	0.04	0.06⋆
UEDIN	-0.11	-0.08	-0.05	-0.04	-	0.02
NICT	-0.13	-0.10	-0.07	-0.06	-0.02	-
score	0.09	0.07	0.03	0.02	-0.02	-0.04
rank	1–6	1–6	1–6	1–6	1–6	1–6

Table 23: Head to head comparison for Turkish→English systems.

	ONLINE-B	UEDIN	ALIBABA-ENS-A	NICT	ALIBABA-ENS-B	ONLINE-G	RWTH	ONLINE-A
ONLINE-B	-	0.05	0.06	0.15⋆	0.17†	0.22‡	0.34‡	0.53‡
UEDIN	-0.05	-	0.01	0.09⋆	0.11⋆	0.16†	0.28‡	0.48‡
ALIBABA-ENS-A	-0.06	-0.01	-	0.09	0.10	0.16†	0.28‡	0.47‡
NICT	-0.15	-0.09	-0.09	-	0.02	0.07	0.19‡	0.38‡
ALIBABA-ENS-B	-0.17	-0.11	-0.10	-0.02	-	0.05	0.17†	0.36‡
ONLINE-G	-0.22	-0.16	-0.16	-0.07	-0.05	-	0.12⋆	0.31‡
RWTH	-0.34	-0.28	-0.28	-0.19	-0.17	-0.12	-	0.19‡
ONLINE-A	-0.53	-0.48	-0.47	-0.38	-0.36	-0.31	-0.19	-
score	0.28	0.22	0.22	0.13	0.11	0.06	-0.06	-0.25
rank	1–6	1–6	1–6	1–6	1–6	1–6	7	8

Table 24: Head to head comparison for English→Turkish systems.

	CUNI-TRANSFORMER	UEDIN	NEWSTEST2018-REF	ONLINE-B	ONLINE-A	ONLINE-G
CUNI-TRANSFORMER	-	0.15‡	0.18‡	0.54‡	0.85‡	1.02‡
UEDIN	-0.15	-	0.04	0.39‡	0.70‡	0.88‡
NEWSTEST2018-REF	-0.18	-0.04	-	0.36‡	0.66‡	0.84‡
ONLINE-B	-0.54	-0.39	-0.36	-	0.31‡	0.48‡
ONLINE-A	-0.85	-0.70	-0.66	-0.31	-	0.18‡
ONLINE-G	-1.02	-0.88	-0.84	-0.48	-0.18	-
score	0.67	0.52	0.48	0.13	-0.18	-0.35
rank	1	2–3	2–3	4	5	6

Table 25: Head to head comparison for Czech→English systems, based on source-based DA.