Findings of the 2017 Conference on Machine Translation (WMT17)

Ondřej Bojar
Charles University

Rajen Chatterjee
FBK

Christian Federmann
Microsoft Research

Yvette Graham
Dublin City University

Barry Haddow
Univ. of Edinburgh

Shujian Huang
Nanjing University

Matthias Huck
LMU Munich

Philipp Koehn
JHU / Edinburgh

Qun Liu
Dublin City University

Varvara Logacheva
MIPT Moscow

Christof Monz
Univ. of Amsterdam

Matteo Negri
FBK

Matt Post
Johns Hopkins Univ.

Raphael Rubino
DFKI & Saarland Univ.

Lucia Specia
Univ. of Sheffield

Marco Turchi
FBK

Abstract

This paper presents the results of the WMT17 shared tasks, which included three machine translation (MT) tasks (news, biomedical, and multimodal), two evaluation tasks (metrics and run-time estimation of MT quality), an automatic post-editing task, a neural MT training task, and a bandit learning task.

1 Introduction

We present the results of the shared tasks of the Second Conference on Statistical Machine Translation (WMT) held at EMNLP 2017. This conference builds on eleven previous editions of WMT as workshops and conference (Koehn and Monz, 2006; Callison-Burch et al., 2007, 2008, 2009, 2010, 2011, 2012; Bojar et al., 2013, 2014, 2015, 2016a).

This year we conducted several official tasks. We report in this paper on three tasks:

- news translation (Section 2, Section 3)
- quality estimation (Section 4)
- automatic post-editing (Section 5)

The conference featured additional shared tasks that are described in separate papers in these proceedings:

- metrics (Bojar et al., 2017a)
- multimodal machine translation and multilingual image description (Elliott et al., 2017)
- biomedical translation (Jimeno Yepes et al., 2017)
- neural MT training (Bojar et al., 2017b)
- bandit learning (Sokolov et al., 2017)

In the news translation task (Section 2), participants were asked to translate a shared test set, optionally restricting themselves to the provided training data (constraint condition). We held 14 translation tasks this year, between English and each of Chinese, Czech, German, Finnish, Latvian, Russian, and Turkish. The Latvian and Chinese translation tasks were new this year. Latvian is a lesser resourced data condition on challenging language pair. Chinese allowed us to co-operate with an ongoing evaluation campaign on Asian languages organized alongside the Chinese Workshop on Machine Translation (CWMT).[1] System outputs for each task were evaluated both automatically and manually.

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 based on judgments of adequacy on a 100-point scale, a method we explored last year with convincing results in terms of the trade-off between annotation effort and reliable distinctions between systems.

The quality estimation task (Section 4) this year included three subtasks: sentence-level prediction of post-editing effort scores, word and phrase-level prediction of good/bad labels. Datasets

[1] `http://nlp.nju.edu.cn/cwmt2017/evaluation.en.html`

Proceedings of the Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 169–214
Copenhagen, Denmark, September 711, 2017.

were released with English→German IT translations and German→English Pharmaceutical translations for all subtasks.

The automatic post-editing task (Section 5) examined automatic methods for correcting errors produced by an unknown machine translation system. Participants were provided with training triples containing source, target and human post-edits, and were asked to return automatic post-edits for unseen (source, target) pairs. In this third round, the task focused on correcting English→German translations in the IT domain and German→English translations in the Pharmaceutical domain.

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 statistical machine translation, automatic evaluation, or prediction of translation quality. News translations are also available for interactive visualization and comparison of differences between systems at `http://wmt.ufal.cz/` using MT-ComparEval (Sudarikov et al., 2016).

2 News Translation Task

The recurring WMT task examines translation between English and other languages in the news domain. As in the previous years, we include German, Czech, Russian, Finnish, and Turkish. New languages this years are Latvian and Chinese.

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 before. We took about 1500 English sentences and translated them into the other 5 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 stories were translated by professional translators, funded by the EU Horizon 2020 projects CRACKER and QT21 (German, Czech, Latvian), by Yandex[3], a Russian search engine company (Turkish, Russian), and by BAULT, a research community on building and using language technology funded by the University of Helsinki (Finnish). 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.

For Latvian, the test set size was 2000 sentences, and an additional 2000 sentences were released as development set.

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[4], Common Crawl, SETIMES2 , Russian-English parallel data provided by Yandex, Wikipedia Headlines provided by CMU) and some were updated (United Nations, CzEng v1.6 (Bojar et al., 2016b), News Commentary v12, monolingual news data). A new corpis is the EU Press Release parallel corpus for German, Finnish, and Latvian.

For Latvian and Chinese a number of new corpora were released. For Latvian this data was prepared by the University of Latvia and Tilde, the Chinese corpora were prepared by the Institutes of Computing Technology and of Automation, Chinese Academy of Science, Northeastern University (China) and Datum Data Co., Ltd.

Some statistics about the training materials are given in Figure 1.

2.3 Submitted systems

We received 103 submissions from 31 institutions. The participating institutions and their entry names are listed in Table 2; each system did not necessarily appear in all translation tasks. We also

[2] `http://statmt.org/wmt17/results.html`

[3] `http://www.yandex.com/`

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

Europarl Parallel Corpus

	German ↔ English		Czech ↔ English		Finnish ↔ English		Latvian ↔ English	
Sentences	1,920,209		646,605		1,926,114		637,599	
Words	50,486,398	53,008,851	14,946,399	17,376,433	37,814,266	52,723,296	11,957,078	15,412,186
Distinct words	381,583	115,966	172,461	63,039	693,963	115,896	289,849	137,244

News Commentary Parallel Corpus

	German ↔ English		Czech ↔ English		Russian ↔ English		Chinese ↔ English	
Sentences	270,769		211,284		222,390		332,525	
Words	6,087,255	5,924,001	4,057,726	4,545,443	4,759,919	5,068,124	–	5,123,145
Distinct words	285,017	181,203	295,447	157,800	317,074	169,315	–	164,103

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

EU Press Release Parallel Corpus

	German ↔ English		Finnish ↔ English		Latvian ↔ English	
Sentences	1,329,041		583,223		306,588	
Words	22,078,112	22,998,930	6,823,630	10,063,161	4,250,672	5,135,993
Distinct words	642,591	347,021	465,355	189,316	200,773	121,401

Latvian Parallel Corpora

	LETA News		Online Books		Corpus of Eu. Parliament	
	Latvian ↔ English		**Latvian ↔ English**		**Latvian ↔ English**	
Sentences	15,671		9,577		3,542,280	
Words	340,394	438,666	63,233	82,665	30,177,230	37,158,634
Distinct words	62,734	41,252	19,191	9,104	604,110	416,932

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 Parallel Corpus

	Czech ↔ English	
Sentences	62,493,539	
Words	611,094,888	688,534,994
Distinct	8,017,713	5,738,815

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 (9), Brisbane Times (1), CBS News (5), CNN (1), Daily Mail (10), Euronews (1), Fox News (2), Globe and Mail (1), Guardian (3), Independent (2), Los Angeles Times (1), Novinte (1), New York Times (8), Reuters (4), Russia Today (3), Scotsman (1), Sydney Morning Herald (4), Telegraph (1), The Local (1), UPI (4)
Chinese	Ifeng (82), People Daily (14), Sina (14), Xinhua (8)
Czech	aktuálně.cz (10), blesk.cz (4), blisty.cz (1), deník.cz (1), iDNES.cz (14), ihned.cz (4), lidovky.cz (8), Novinky.cz (5), Reflex (1), tyden.cz (4), ZDN (2)
German	Abendzeitung München (1), Abendzeitung Nürnberg (1), ARD (1), Augsburger Allgemeine (1), Bergedorfer Zeitung (1), Braunschweiger Zeitung (1), Der Standard (2), Deutsche Welle (1), Dülmener Zeitung (1), Euronews (1), Frankfurter Rundschau (2), Generalanzeiger Bonn (1), Göttinger Tageblatt (1), Handelsblatt (4), In Franken (4), In Südthüringen (1), Kieler Nachrichten (2), Kreisanzeiger (1), Kreiszeitung (3), Krone (1), Kölner Stadt Anzeiger (2), Merkur (1), Morgenpost (3), Neue Presse Coburg (1), Nordbayerischer Kurier (1), oe24 (1), Potzdamer Neueste Nachrichten (1), Passauer Neue Presse (1), Pforzheimer Zeitung (1), Rheinzeitung (1), Rundschau (1), Schwarzwälder Bote (2), Südkurier (1), Süddeutsche Zeitung (1), Usinger Anzeiger (1), Westfälischer Anzeiger (1), Westfälische Nachrichten (3), Westdeutsche Zeitung (4), Zeit (1), Waiblinger Kreiszeitung (4).
Finnish	Etelä-Saimaa (2), Etelä-Suomen Sanomat (1), Helsingin Sanomat (14), Ilkka (10), Iltalehti (16), Ilta-Sanomat (16), Kaleva (9), Kansan Uutiset (3), Karjalainen (10), Kouvolan Sanomat (2), Loimaan Lehti (1).
Latvian	Dienas Bizness (3), Delfi (11), Diena (13), grenet.lv (1), LSM (10), NRA (9), Talsu Vestis (1), TV Net (21)
Russian	aif (), dp.ru (2), eg-online.ru (2), gazeta.ru (5), gzt-sv.ru (1), Izvestiya (7), Kommersant (16), Lenta (17), lgng (5), MK RU (4), nov-pravda.ru (1), Novaya Gazeta (3), pnp.ru (4), rg.ru (1), rusplit.ru (1), Vedomosti (1), Versia (2), Vesti (3), VM News (1), zr.ru (3)
Turkish	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.

Europarl Language Model Data

	English	German	Czech	Finnish	Latvian
Sentences	2,218,201	2,176,537	668,595	2,120,739	667, 241
Words	59,848,044	53,534,167	14,946,399	39,511,068	12,092,389
Distinct words	123,059	394,781	172,461	711,868	160,312

News Language Model Data

	English	German	Czech	Russian	Finnish
Sentences	166,127,560	221,793,141	59,184,372	31,285,072	10,938,701
Words	3,816,723,867	3,938,344,482	974,167,234	572,672,132	137,162,922
Distinct words	5,895,731	17,824,672	4,011,712	2,929,646	,3557,784

Common Crawl Language Model Data

	English	German	Czech	Russian	Finnish	Romanian	Turkish
Sent.	3,074,921,453	2,872,785,485	333,498,145	1,168,529,851	157,264,161	288,806,234	511,196,951
Words	65,128,419,540	65,154,042,103	6,694,811,063	23,313,060,950	2,935,402,545	8,140,378,873	11,882,126,872
Dist.	342,760,462	339,983,035	50,162,437	101,436,673	47,083,545	37,846,546	88,463,295

Test Set

	Czech ↔ EN		German ↔ EN		Finnish ↔ EN		Latvian ↔ EN	
Sentences.	3,005		3,004		3,002		2,001	
Words	54,630	61,958	60,963	64,760	45,472	62,769	39,064	47,832
Distinct words	14,462	8,544	12,514	8,997	16,156	8,552	11,708	7,435

	Russian ↔ EN		Turkish ↔ EN		Chinese ↔ EN	
Sentences.	3,001		3,007		2,001	
Words	59,912	69,847	55,303	67,927	–	54,011
Distinct words	17,391	9,386	14,864	8,664	–	7,710

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.

ID	Institution
AALTO	Aalto University(Grönroos et al., 2017)
AFRL-MITLL	Air Force Research Lab / MIT Lincoln Lab (Gwinnup et al., 2017)
APERTIUM	Apertium / Helsinki University (Hurskainen and Tiedemann, 2017)
C-3MA	Tartu-Riga-Zürich (Rikters et al., 2017)
CASICT-DCU	Chinese Academy of Sciences / Dublin City University (Zhang et al., 2017)
CU-CHIMERA	Charles University (Sudarikov et al., 2017)
FBK	Fondazione Bruno Kessler (Di Gangi et al., 2017)
HUNTER	Hunter College, City University of New York (Xu et al., 2017)
HY	Helsinki University (Östling et al., 2017)
JAIST	Japan Advanced Institute of Science and Technology (Trieu et al., 2017)
JHU	Johns Hopkins University (Ding et al., 2017)
KIT	Karlsruhe Institute of Technology (Pham et al., 2017)
LIMSI	LIMSI (Burlot et al., 2017)
LIUM-CVC	University of Le Mans / Universitat Autonoma de Barcelona (García-Martínez et al., 2017)
LMU	LMU Munich (Huck et al., 2017)
NMT-AVE-MULTI-CS	
NRC	National Research Council, Canada
OREGON	Orgon State University
PJATK	Polish-Japanese Academy of Information (Wolk and Marasek, 2017)
PROMT	PROMT Rule-Based System
QT21	QT21 project system combination (Peter et al., 2017b)
ROCMT	University of Rochester (Holtz et al., 2017)
RWTH	RWTH Aachen (Peter et al., 2017a)
SOGOU	Sogou Inc. (Wang et al., 2017)
SYSTRAN	Systran (Deng et al., 2017)
TALP-UPC	TALP, Technical University of Catalonia (Escolano et al., 2017)
TILDE	Tilde (Pinnis et al., 2017)
UEDIN	University of Edinburgh (Sennrich et al., 2017)
USFD	University of Sheffield
UU	Uppsala University
XMU	Xiamen University (Tan et al., 2017b)

Table 2: Participants in the shared translation task. Not all teams participated in all language pairs. The translations from the commercial and 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.

included 39 online statistical MT systems (originating from 4 services), which we anonymized as ONLINE-A,B,F,G.

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, these online and commercial systems are treated as unconstrained during the automatic and human evaluations.

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 is carried out this year.

In previous years, we asked human annotators to rank the outputs of five systems. From these rankings, we produced pairwise translation comparisons, and applied the TrueSkill algorithm (Herbrich et al., 2007; Sakaguchi et al., 2014) to produce system rankings. We refer to this approach as the *relative ranking* (RR) approach, so named because the pairwise comparisons denote only relative ability between a pair of systems, and cannot be used to infer absolute quality. For example, RR can be used to discover which systems perform better than others, but RR does not provide any information about the absolute quality of system translations, i.e. it provides no information about how far a given system is from producing perfect output according to a human user.

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 last year's evaluation campaign included parallel assessment of a subset of News task language pairs evaluated with 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 last year (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). This year, we thus employ DA only. 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 with a new implementation of DA. Researcher involvement is needed particularly for translations out-of-English.

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.[5] 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 and details on time spent in Section 3.6 below.

3.1 Evaluation Campaign Overview

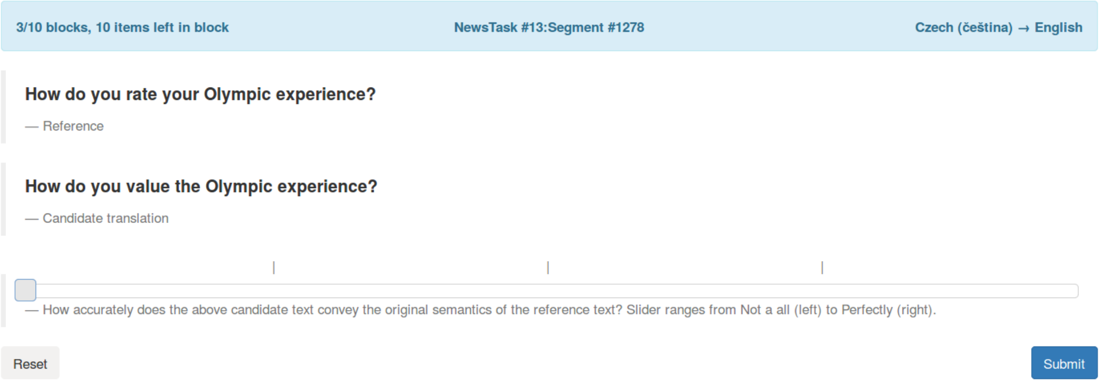
In terms of the News translation task manual evaluation, a total of 151 individual researcher accounts were involved, and 754 turker accounts.[6] Researchers in the manual evaluation came from 29 different research groups and contributed judgments of 125,693 translations, while 237,200 translation assessment scores were submitted in total by the crowd.[7]

Under ordinary circumstances, each assessed translation would correspond to a single individual scored segment. However, since many systems

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

[6] Numbers do not include the 954 workers on Mechanical Turk who did not pass quality control.

[7] Numbers include quality control items for workers who passed quality control but omit the additional 151,200 assessments collected on Mechanical Turk where a worker did not pass quality control.



3/10 blocks, 10 items left in block

NewsTask #13:Segment #1278

Czech (čeština) → English

How do you rate your Olympic experience?

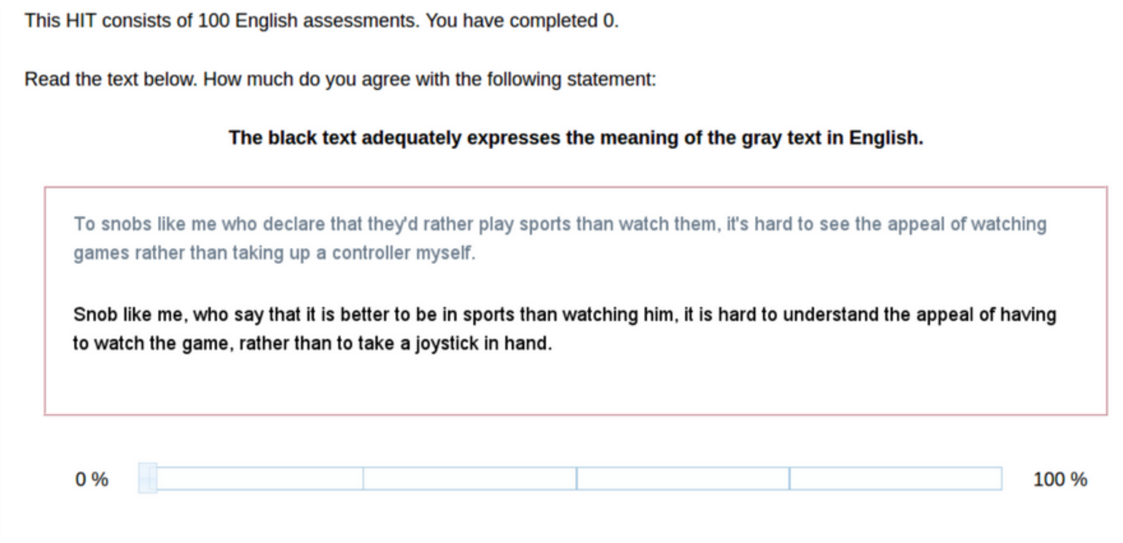
— Reference

How do you value the Olympic experience?

— Candidate translation

— How accurately does the above candidate text convey the original semantics of the reference text? Slider ranges from Not a all (left) to Perfectly (right).

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.



This HIT consists of 100 English assessments. You have completed 0.

Read the text below. How much do you agree with the following statement:

The black text adequately expresses the meaning of the gray text in English.

To snobs like me who declare that they'd rather play sports than watch them, it's hard to see the appeal of watching games rather than taking up a controller myself.

Snob like me, who say that it is better to be in sports than watching him, it is hard to understand the appeal of having to watch the game, rather than to take a joystick in hand.

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

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. This year we only combine human assessments in this way if the string of text belonging to multiple systems is exactly identical. For example, even small differences in punctuation disqualify the potential combination of similar system outputs into a single human assessment, and this is due to lack of evidence about what kinds of minor differences might impact human evaluation.

Table 3 shows the numbers of segments for which distinct MT systems participating in the News task produced identical outputs. English to Czech is the only language pair to include systems that do not belong to the news task, the additional NMT Training task systems, and we include a breakdown of duplicate translations by each task for that language pair in Table 3. The biggest saving in terms of exact duplicate translations for multiple systems was made in the News task for English to German.

3.2 Data Collection

The system ranking is produced from a large set of human assessments, each of which indicates the absolute quality of the output of a system. Annotations are collected in an evaluation campaign that enlists participants in the shared task to help. Each team is asked to contribute 8 hours anno-

Language Pair	# Systems	# Segs	# Total Segs	# Unique Segs	Overall Saving
Chinese→English	16	2,001	32,016	30,772	3.9 %
Czech→English	4	3,005	12,020	11,501	4.3 %
German→English	11	3,004	33,044	29,513	10.7 %
Finnish→English	6	3,002	18,012	17,766	1.4 %
Latvian→English	9	2,001	18,009	17,441	3.2 %
Russian→English	9	3,001	27,009	25,430	5.8 %
Turkish→English	10	3,007	30,070	28,672	4.6 %
English→Chinese	11	2,001	22,011	21,626	1.7 %
English→Czech	14	3,005	42,070	37,774	10.2 %
News	8	3,005	24,040	21,261	11.6 %
NMT Training	6	3,005	18,030	17,098	5.2 %
English→German	16	3,004	48,064	41,918	12.8 %
English→Finnish	12	3,002	36,024	34,688	3.7 %
English→Latvian	17	2,001	34,017	30,928	9.1 %
English→Russian	9	3,001	27,009	25,807	4.5 %
English→Turkish	8	3,007	24,056	23,540	2.1 %

Table 3: Total segments prior to sampling for manual evaluation and savings made by combining identical segments (Segs) produced by multiple MT systems in the News (all language pairs) and NMT Training task (English→Czech only).

tation time, which we estimated this year at 16 100-translation HITs per primary system submitted. We continue to use the open-source Appraise[8] (Federmann, 2012) tool for our data collection, in addition to Amazon Mechanical Turk.[9] Table 4 shows total numbers of human assessments collected in WMT17 contributing to final scores for systems.

When summarizing and comparing annotation times recorded on Appraise and Mechanical Turk, both encounter possible challenges in terms of idle times exaggerating summary statistics. We explore this issue in detail in Section 3.6, and for the summary that follows, assessment times for Appraise that appear to include very lengthy idle times are each replaced with a realistic average time per assessment, as described in Section 3.6. In total, our human annotators spent nearly 24 days and 22 hours working on Appraise, and 47 days and 23 hours annotating via crowdsourcing.[10] This gives an average annotation time of 4 hours per researcher using Appraise and 1 hour 32 minutes contribution by individual workers on Mechanical Turk.[11] Compared to last year's RR evaluation, we see a reduction in average time commitment per researcher, which was 6.4 hours in WMT16.

In this year's evaluation, since it is the first time DA has been used with non-crowdsourced human evaluators, estimates of expected assessment completion times were used to guess the required time commitment by each participating team. Similar to the previous campaigns, several of the Appraise annotators passed the mark of required numbers of annotations (the maximum number being 5,240 translation assessments) with the most patient annotator contributing close to 22.5 hours of work. However, for one language pair, English to Latvian, insufficient annotations were contributed by researchers, which we suspect was caused by the difficulty in sourcing Latvian speakers.

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

[8] `https://github.com/cfedermann/Appraise`

[9] `https://www.mturk.com`

[10] Numbers do not include the 2,106,918 seconds of annotation provided by workers who did not pass quality control.

[11] Times for Mechanical Turk workers do not include workers who failed to pass quality control checks. Some but not all of the HITs that do not pass quality control checks are rejected and therefore go unpaid. A portion of unusable data is accepted and paid due to the possibility that some diligent workers may simply lack the required literary skills to pass quality control.

Language Pair	**Systems**	**Comps**	**Comps/Sys**	**Assessments**	**Assess/Sys**
Chinese→English	16	–	–	38,736	2,421
Czech→English	4	–	–	21,992	5,498
German→English	11	–	–	36,189	3,290
Finnish→English	6	–	–	27,545	4,591
Latvian→English	9	–	–	30,321	3,369
Russian→English	9	–	–	24,837	2,760
Turkish→English	10	–	–	25,853	2,585
English→Chinese	11	–	–	16,253	1,478
English→Czech	15	–	–	32,564	2,171
English→German	16	–	–	10,229	639
English→Finnish	12	–	–	8,289	691
English→Latvian	17	–	–	6,882	405
English→Russian	9	–	–	25,798	2,866
English→Turkish	8	–	–	2,219	277
Total Researcher	153	–	–	107,902	705
Total Crowd	85	–	–	199,805	2,351
Total WMT17	**153**	–	–	**307,707**	**2,011**
WMT16	138	569,287	4,125.2	284,644	2,062
WMT15	131	542,732	4,143.0	271,366	2,071
WMT14	110	328,830	2,989.3	164,415	1,494
WMT13	148	942,840	6,370.5	471,420	3,185
WMT12	103	101,969	999.6	50,985	495
WMT11	133	63,045	474.0	31,522	237

Table 4: Amount of data (assessments after removal of quality control items and "de-collapsing" *multi-system outputs*) collected in the WMT17 manual evaluation campaign. The final six rows report summary information from previous years of the workshop. Note how many rankings we get for Czech language pairs; these include systems from the NMT Training shared task.

3.3 Crowd Quality Control

Translations are arranged in sets of 100-translation HITs as this allows a minimum number of pairs of quality control translations to be collected from each worker who participates, while at the same time allowing sufficient separation of assessment of quality control translation pairs so that human assessors are highly unlikely to simply remember the score they assigned to the initial assessed translation. Details of the three kinds of quality control translation pairs employed by DA are provided in Table 5: we repeat pairs (expecting a similar judgement), damage MT outputs (expecting significantly worse scores) and use references instead of MT outputs (expecting high scores). Bad reference pairs are 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 in itself will comprise a fluent sequence of words (making it difficult to tell that the sentence is low quality without reading the entire sentence) while at the same time making its presence highly likely to sufficiently change the meaning of the MT output so that it causes a noticeable degradation. The length of the phrase to be replaced is determined by the number of words in the translation to be degraded, as follows:

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

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: Quality control translation pairs hidden within 100-translation HITs, numbers of items are provided in parentheses.

In total, 60 items in a 100-translation HIT serve in quality control checks but 40 of those are regular judgements 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%.

3.4 Annotator Agreement

When an analogue (or 0–100 point, in practice) scale is employed, agreement cannot be measured using the conventional Kappa coefficient, ordinarily applied to evaluation of human assessment where judgments are discrete categories or preferences. Instead, we filter crowd-sourced human assessors by how consistently they rate translations of known distinct quality using 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 and even robotic HITs. We therefore employ DA's quality control mechanism, facilitated by the use of DA's analogue rating scale.

Assessments belonging to a given crowd-sourced worker who has not demonstrated that they can reliably score bad reference translations significantly lower than corresponding genuine system output translations are filtered out. The p-value produced in a paired significance test of bad reference pair score distributions is used as an estimate of human assessor reliability. Assessments of workers whose p-value does not fall below the conventional 0.05 threshold are omitted from the evaluation of systems, since they do not reliably score degraded translations lower than corresponding MT output translations.

Table 6 shows the number of unique workers who evaluated MT output on Mechanical Turk via DA, those who met our filtering requirement by showing a significantly lower score for bad reference items, and the proportion of those workers who simultaneously showed no significant difference between scores they attributed in repeat assessment of identical translations. 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 of 97–100%.

We also see in Table 6 that the number of excluded Mechanical Turk workers can be high for many languages, between 42 and 58% for English HITs, 72% for Russian and 81% for Chinese. The variance in English annotations for different source languages are consistent with previous DA evaluations and we do not believe this is caused in any significant way by the source language. With respect to the choice of target language, however, in general DA evaluation for languages with fewer speakers on Mechanical Turk, such as Russian and Chinese, do tend to encounter higher rates of gaming. Since HITs are slower to complete, due to fewer workers with that language, HITs are live for a longer duration on the service and gamer-type workers have a greater opportunity to attempt payment for them.

This year, bad reference items were only collected for crowd-sourced assessments. For information on quality control statistics for non crowd-sourced workers see this year's human evaluation of the APE task, Section 5.5, where student volunteers were employed and although only 11 annotators were involved in total, 100% of those passed DA's quality control filter.

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

	All	(A) Sig. Diff. Bad Ref.	(A) & No Sig. Diff. Exact Rep.
Czech→English	154	89 (58%)	87 (98%)
German→English	398	201 (51%)	194 (97%)
Finnish→English	264	106 (40%)	102 (96%)
Latvian→English	332	123 (37%)	122 (99%)
Russian→English	274	148 (54%)	144 (97%)
Turkish→English	344	107 (31%)	103 (96%)
Chinese→English	386	161 (42%)	158 (98%)
English→Russian	82	23 (28%)	23 (100%)
English→Chinese	43	8 (19%)	8 (100%)
Total	**1708**	**754 (44%)**	**733 (97%)**

Table 6: Number of unique Mechanical Turk workers, (A) those whose scores for bad reference pairs were significantly different and numbers of unique human assessors in (A) whose scores for exact repeat assessments also showed no significant difference.

system is computed as the average of its segment scores (Ave z in Table 7). Results are also reported for average scores for systems, computed in the same way but without any score standardization applied (Ave % in Table 7).

Table 7 includes final DA scores for all systems participating in WMT17 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. Figure 5 shows the underlying head-to-head significance test results for all pairs of systems.

3.6 Crowd versus Researcher Results Comparison

Finally, although we have combined all data collected via crowd-sourcing and researchers to produce the overall results of the shared task, sufficient assessments were collected to produce system scores independently in both set-ups for three language pairs. Table 8 shows the Pearson correlation between DA scores for systems when evaluated by researchers with scores produced via crowd-sourcing, showing high levels of agreement reached overall for all language pairs as correlations range from 0.98 to 0.997.

In terms on annotation times, some differences in the way HIT durations are recorded within Appraise and Mechanical Turk make a comparison of annotation times for researchers and crowd-sourced workers not entirely straightforward. On the one hand, it is possible for a Mechanical Turk (Mturk) worker, attempting to game the system, to leave the window idle in order to obscure a lack of effort, while on Appraise, researcher annotation times will naturally include idle times due to interruptions of some kind.

The degree to which annotation times can be exaggerated for Mturk workers is quite limited, however. Firstly, since we impose quality control checks throughout Mturk HITs, it won't be possible for many workers to meet the quality threshold without genuinely spending a minimum amount of time on assessments. Additionally, we impose a hard time limit of 90 minutes duration to each 100-translation HIT on Mturk (this corresponds to an average maximum completion time of 54 seconds per translation) which limits the amount of exaggeration of completion times that can take place. The situation on Appraise is quite different however, and idle times could potentially severely skew annotation time analysis.

Figure 6(a) shows annotation times recorded for our HITs on Mechanical Turk and Figure 6(b) shows equivalent times for Appraise, where both sets of completion times have been sorted from shortest to longest duration. Examining the y-axis of the Appraise plot in Figure 6(b) shows the maximum completion time for a single translation to be at a whopping 329,578 seconds (3.8 days), revealing the extent to which the inclusion of idle times for Appraise runs the risk of exaggerating annotation times for researchers, while on Mechanical Turk, Figure 6(a), the 90 minute HIT du-

Chinese→English

#	Ave %	Ave z	System
1	73.2	0.209	SogouKnowing-nmt
	73.8	0.208	uedin-nmt
	72.3	0.184	xmunmt
4	69.9	0.113	online-B
	70.4	0.109	online-A
	69.8	0.079	NRC
7	67.9	0.023	jhu-nmt
	66.9	−0.016	afrl-mitll-opennmt
	67.1	−0.026	CASICT-DCU_NMT
	65.4	−0.058	ROCMT
11	64.3	−0.107	Oregon-State-Uni-S
12	61.7	−0.209	PROMT-SMT
	61.2	−0.265	NMT-Ave-Multi-Cs
	60.0	−0.276	UU-HNMT
	59.6	−0.279	online-F
	59.3	−0.305	online-G

English→Chinese

#	Ave %	Ave z	system
1	73.2	0.208	SogouKnowing-nmt
	72.5	0.178	uedin-nmt
	72.0	0.165	xmunmt
4	69.8	0.065	online-B
	69.5	0.056	jhu-nmt
	68.5	0.035	CASICT-DCU_NMT
	68.2	0.010	online-A
8	64.8	−0.111	Oregon-State-Uni-S
9	59.2	−0.300	UU-HNMT
10	55.9	−0.438	online-G
11	53.1	−0.504	online-F

Czech→English

#	Ave %	Ave z	system
1	74.6	0.181	uedin-nmt
2	71.9	0.068	online-B
3	68.3	−0.068	online-A
4	62.7	−0.268	PJATK

English→Czech

#	Ave %	Ave z	system
1	62.0	0.308	uedin-nmt
2	59.7	0.240	online-B
3	55.9	0.111	limsi-factored-norm
	55.2	0.102	LIUM-FNMT
	55.2	0.090	LIUM-NMT
	54.1	0.050	CU-Chimera
	53.3	0.029	online-A
8	44.9	−0.236	TT-ufal-8GB
9	42.2	−0.315	TT-afrl-4GB
	41.9	−0.327	PJATK
	40.7	−0.373	TT-base-8GB
	40.5	−0.376	TT-afrl-8GB
13	36.5	−0.486	TT-ufal-4GB
	36.6	−0.493	TT-denisov-4GB

German→English

#	Ave %	Ave z	system
1	78.2	0.213	online-B
	76.6	0.169	online-A
	76.6	0.165	KIT
	76.6	0.162	uedin-nmt
	75.8	0.131	RWTH-nmt-ensemb
	74.5	0.098	SYSTRAN
7	72.9	0.029	LIUM-NMT
8	70.2	−0.058	TALP-UPC
	69.8	−0.072	online-G
	68.6	−0.103	C-3MA
11	64.1	−0.260	online-F

English→German

#	Ave %	Ave z	system
1	72.9	0.257	LMU-nmt-reranked
2	70.2	0.158	online-B
	69.8	0.139	uedin-nmt
	68.9	0.092	SYSTRAN
	66.9	0.035	LMU-nmt-single
	66.7	0.022	KIT
	66.4	0.015	xmu
	66.6	0.006	LIUM-NMT
	66.0	−0.003	RWTH-nmt-ensemb
10	60.1	−0.233	online-A
	60.3	−0.234	PROMT-Rule-based
	58.9	−0.270	C-3MA
	58.1	−0.301	fbk-nmt-comb
	55.2	−0.391	TALP-UPC
	54.9	−0.440	online-F
	53.2	−0.491	online-G

Finnish→English

#	Ave %	Ave z	system
1	73.8	0.407	online-B
2	67.5	0.220	online-G
3	62.6	0.041	online-A
4	58.8	−0.095	TALP-UPC
5	52.1	−0.316	Hunter-MT
6	44.6	−0.559	apertium

English→Finnish

#	Ave %	Ave z	system
1	59.6	0.378	online-B
	57.8	0.305	HY-HNMT
3	51.6	0.090	online-G
	51.3	0.060	jhu-nmt-latt-resc
	49.3	−0.004	AaltoHnmtMultitask
6	46.4	−0.102	AaltoHnmtFlatcat
	46.7	−0.109	online-A
	45.8	−0.115	HY-SMT
	43.5	−0.192	HY-AH
	43.4	−0.204	jhu-pbmt
11	40.8	−0.298	TALP-UPC
12	8.0	−1.428	apertium

Latvian→English

#	Ave %	Ave z	system
1	76.2	0.266	online-B
	76.2	0.245	tilde-nc-nmt-smt
3	71.4	0.087	uedin-nmt
	71.0	0.083	tilde-c-nmt-smt
5	67.3	−0.039	online-A
6	64.4	−0.137	jhu-pbmt
7	63.4	−0.187	C-3MA
	62.2	−0.199	Hunter-MT
9	56.3	−0.436	PJATK

English→Latvian

#	Ave %	Ave z	system
1	54.4	0.196	tilde-nc-nmt-smt
	51.6	0.121	online-B
	51.1	0.104	tilde-c-nmt-smt
	50.8	0.075	limsi-fact-norm
	50.0	0.058	usfd-cons-qt21
	47.1	−0.014	QT21-Comb
	47.3	−0.027	usfd-cons-kit
	45.7	−0.063	KIT
	45.2	−0.072	uedin-nmt
	44.9	−0.099	tilde-nc-smt
	43.2	−0.157	LIUM-FNMT
	43.0	−0.198	LIUM-NMT
	40.1	−0.253	HY-HNMT
	37.5	−0.341	online-A
	36.1	−0.368	jhu-pbmt
	33.3	−0.457	C-3MA
17	18.8	−0.947	PJATK

Russian→English

#	Ave %	Ave z	system
1	82.0	0.271	online-B
2	77.6	0.126	online-G
3	76.5	0.081	NRC
	76.1	0.057	online-A
	74.9	0.017	afrl-mitll-comb
	74.6	0.005	afrl-mitll-opennmt
	74.2	0.002	uedin-nmt
	74.7	−0.011	jhu-pbmt
9	65.9	−0.288	online-F

English→Russian

#	Ave %	Ave z	system
1	75.4	0.402	online-B
2	68.2	0.166	uedin-nmt
3	66.5	0.105	online-H
4	65.9	0.080	PROMT-Rule-based
	65.2	0.061	online-A
	65.2	0.054	online-G
7	62.6	−0.018	jhu-pbmt
8	57.3	−0.194	afrl-mitll-backtra
9	46.5	−0.568	online-F

Turkish→English

#	Ave %	Ave z	system
1	68.8	0.294	online-B
	68.5	0.282	online-A
3	61.1	0.050	uedin-nmt
4	58.6	−0.029	online-G
	58.0	−0.083	afrl-mitll-m2w
	57.0	−0.093	afrl-mitll-comb
	56.7	−0.097	LIUM-NMT
8	53.5	−0.183	PROMT-SMT
9	46.4	−0.436	jhu-pbmt
	45.5	−0.475	JAIST

English→Turkish

#	Ave %	Ave z	system
1	53.4	0.513	online-B
2	44.0	0.206	uedin-nmt
3	39.1	0.071	online-A
	35.5	−0.032	online-G
5	32.2	−0.129	LIUM-NMT
6	18.0	−0.554	jhu-nmt-latt-resc
	16.7	−0.597	jhu-pbmt
	15.7	−0.602	JAIST

Table 7: Official results of WMT17 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 $p \leq 0.05$. Systems with gray background indicate use of resources that fall outside the constraints provided for the shared task.

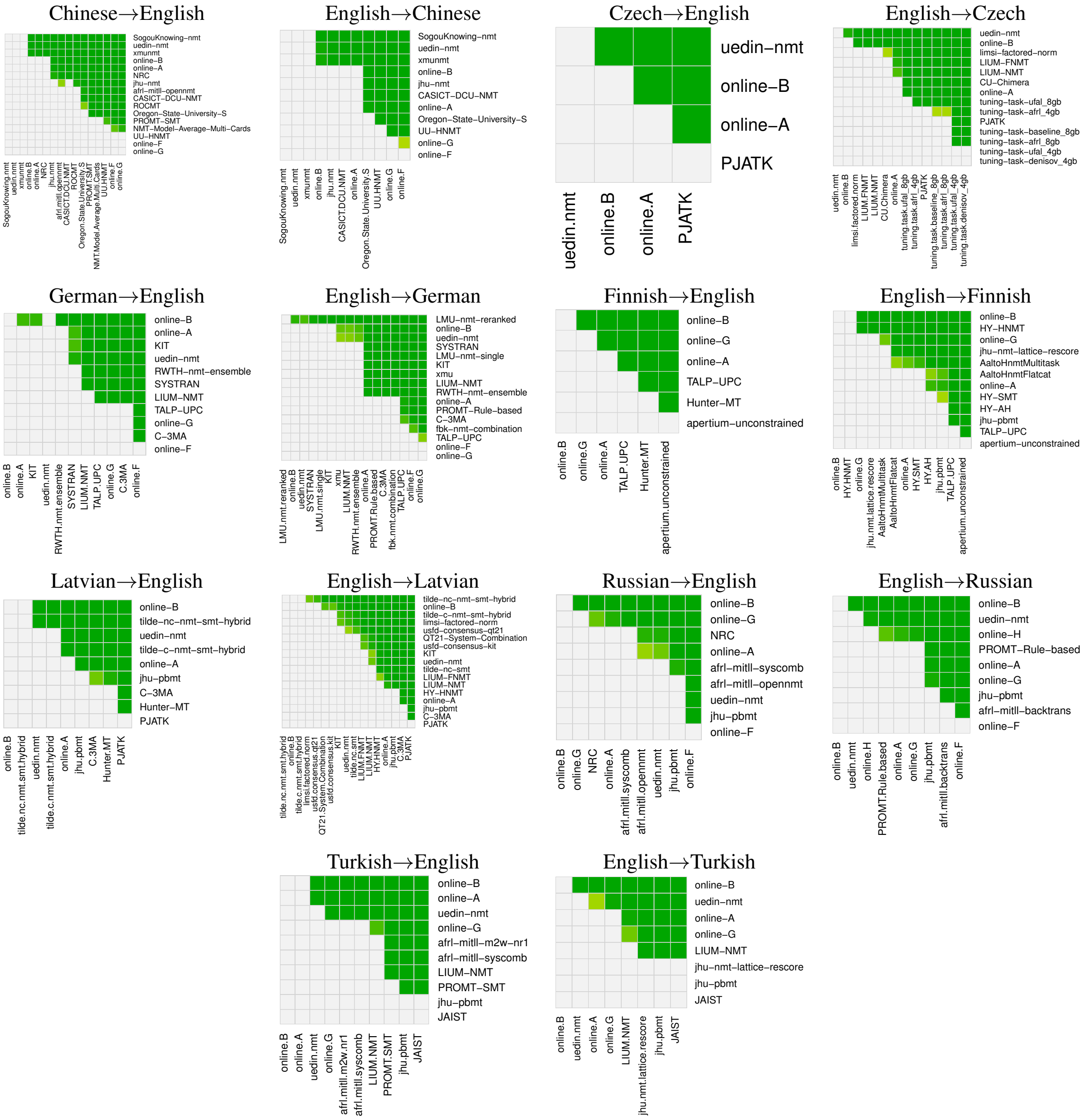


Figure 5: Wilcoxon rank-sum significance test results for pairs of systems competing in the News translation task, where a green cell denotes a significant win for the system in a given row over the system in a given column, at $p \leq 0.05$.

	r	# Researcher	# Crowd
Czech→English	0.997	2,915	2,445
Finnish→English	0.996	1,261	3,245
English→Russian	0.980	867	1,889

Table 8: Pearson correlation (r) between overall DA standardized mean adequacy scores collected via crowd-sourcing (Mturk) and from researchers participating in the shared task (Appraise), numbers of assessments per system (#) are also provided for each set-up.

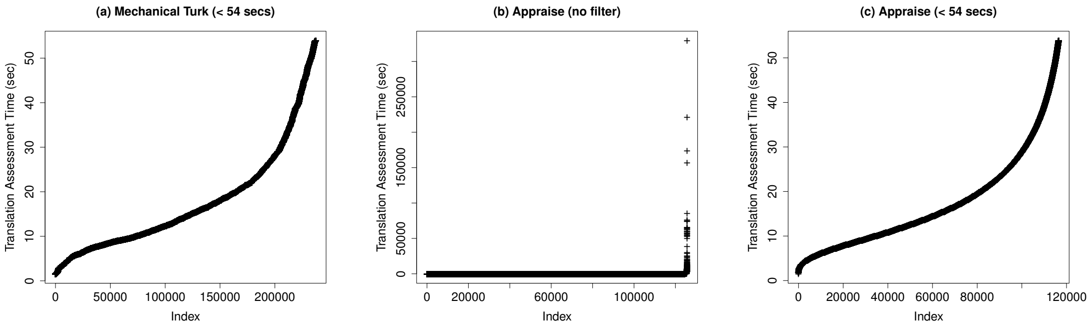


Figure 6: Comparison of Completion Times for (a) Mechanical Turk Assessments; (b) Appraise Assessments (unfiltered); (c) Appraise Assessments (with reasonable cut-off imposed)

ration constraint we impose means that the maximum annotation time per translation is just under 54 seconds.

Although it is possible that assessment times for Mechanical Turk HITs in Figure 6(a) still contain a degree of idle-time exaggeration themselves, the extent to which they could possibly obscure assessment times is vastly less than that of Appraise. Prior to analysis of assessment times, we therefore impose a reasonable limit on what could be considered a realistic maximum annotation time for assessment of a single translation with DA on Appraise. Just to remind ourselves, the assessment of a single translation on Appraise includes: (i) reading a reference translation; (ii) reading the MT output; (iii) considering how well the latter expresses the meaning of the former; (iv) assigning a score via the analogue rating scale; (v) pressing the submit button. We apply the same maximum cut-off applied within Mturk assessments of 54 seconds per translation assessment to Appraise annotation times analysis therefore, which is a reasonable maximum duration for a single translation assessment. Figure 6(c) shows a plot of sorted assessment times for Appraise assessments when this cut-off is applied.

Once overly lengthy idle times have been omitted, it is possible to compare the speed at which researchers and crowd-sourced workers complete DA assessments, in addition to comparing annotation times in this year's DA evaluation with WMT16's RR evaluation as both were completed by researchers. Table 9 shows average annotation times for each human annotator type, and annotation scheme. Annotation times for DA in terms of the average time taken to assess a single translation are straightforward to compute, since a single translation is assessed per screen. Each RR assessment is made up of a relative ranking of five MT output translations, however. Therefore to compute average annotation times for a single translation with RR we simply divide the average time to evaluate five translations by five.

	DA	DA	RR
	Crowd	Researcher	Researcher
WMT16	19.6	–	20.8
WMT17	17.5	17.1	–

Table 9: Average annotation time per translation (in seconds)

Before comparing annotation times, it is important to note that we must take care comparing annotations times collected in two different year's evaluation campaigns, as for researchers, the annotators involved in the evaluation will have some overlap, this is less likely for crowd-sourced workers and in both cases the data involved comes from two different data sets. The evaluation produced by researchers in WMT16 and WMT17 does, however, provide the first data enabling a comparison of annotation speeds for researchers employing DA and RR. Annotation times analysis should only provide an approximate indication of speeds as opposed to tried and tested findings, however, which we hope to provide in the future.

Table 9 shows the reduction in average annotation time resulting from DA's simpler assessment set-up for researchers, from 20.8 seconds per assessment with RR to 17.1 seconds with DA, an approximate reduction of 18%.

Comparing annotation speeds for crowd-sourced workers evaluating with DA in both WMT16 and WMT17, we also see a slight speed up from 19.6 to 17.5 seconds. It is difficult to conclude from a comparison of crowd-sourced workers that this as a genuine speed up as it is likely due at least in part to variance in annotation styles of two different groups of workers drawn from a very large crowd. For example, average annotation times of crowd-sourced workers in the APE task this year was 13.6 seconds with DA where a distinct set of workers was also employed.

In terms of researchers versus crowd-sourced workers evaluating with DA, when we compare this year's results, researchers appear to be marginally quicker, on average approximately 0.4 seconds faster per translation assessment. Although again, this comparison includes average annotation times of crowd-sourced workers that can naturally vary from one group to the next.

Finally, we include a brief comparison in terms of projected time commitments required by participants in future evaluations when the methodology employed is DA rather than RR. In subsequent evaluations, since we have verified that DA results produced by quality controlled crowd-sourcing correspond very closely to researcher results, it should be possible to collect all to-English evaluations via crowd-sourcing. This means that the switch to DA may result in only requiring participants to make a time commitment in terms of out-of-English language pairs. For some research groups this will cut the required manual evaluation time commitment in half.

Assuming a similar number of language pairs as in WMT17 (14 language pairs), an RR manual evaluation, which in previous years required manual evaluation of 100 HITs (each containing 15 translations), amounts to a commitment of assessment of 1,500 translations per submitted system. Considering researchers took on average 20.8 seconds per translation, a team wishing to participate in all language pairs would require a total time commitment of approximately (1,500 x 20.8 seconds x 14 = 436,800 seconds) 121.3 hours. In comparison for DA, even if we stick with the same number of translations per submission (1,500), when we take into account the fact that all of the to-English language pairs can be crowd-sourced as well as the quicker annotation time for DA, the time commitment for such a team would be reduced by approximately 60% to (1,500 x 17.1 seconds x 7 = 179,550 seconds) 49.9 hours.

4 Quality Estimation Task

This shared task builds on its previous five editions to further examine automatic methods for estimating the quality of machine translation output at run-time, without the use of reference translations. It includes the (sub)tasks of word-level, phrase-level and sentence-level estimation. In addition to advancing the state of the art at all prediction levels, our goals include:

- To test the effectiveness of larger (domain-specific and professionally annotated) datasets. We do so by significantly increasing the size of one of last year's training sets.
- To study the effect of language direction and domain. We do so by providing two datasets created in similar ways, but for different domains and language directions.
- To investigate the utility of detailed information logged during post-editing. We do so by providing a score for perceived post-editing effort, post-editing time, keystrokes, and actual edits.
- To measure progress over years at all prediction levels. We do so by using last year's test set for comparative experiments.

This year's shared task provides new training and test datasets for all tasks, and allows participants to explore any additional data and resources deemed relevant. All tasks make use of a large dataset produced from post-editions by professional translators. The data is domain-specific (IT and Pharmaceutical domains) and substantially larger than in previous years. An in-house, in-domain SMT system was used to produce translations for all tasks. System-internal information was made available under request. The data is publicly available but since it was provided by industry collaborators it is subject to specific terms and conditions. However, these have no practical implications on the use of this data for research purposes.

The three tasks are defined as follows: Task 1 at sentence level (Section 4.4), Task 2 at word level (Section 4.5), and Task 3 at phrase level (Section 4.6). Two datasets are used for all tasks (Section 4.3): English-German and German-English SMT translations labelled with task-specific labels. Participants were also provided with a baseline set of features for each task, and a software package to extract these and other quality estimation features and perform model learning (Section 4.1). Participants (Section 4.2) could submit up to two systems for each task. A discussion on the main goals and findings from this year's task is given in Section 4.7.

4.1 Baseline systems

Sentence-level baseline system: For Task 1, QuEst++[12] (2015) was used to extract 17 MT system-independent features from the source and translation (target) files and parallel corpora:

- Number of tokens in the source and target sentences.
- Average source token length.
- Average number of occurrences of the target word within the target sentence.
- Number of punctuation marks in source and target sentences.
- Language model (LM) probability of source and target sentences based on models built using the source or target sides of the parallel corpus used to train the SMT system.
- Average number of translations per source word in the sentence as given by the IBM model 1 extracted using the SMT parallel corpus, and thresholded such that $P(t|s) > 0.2$ or $P(t|s) > 0.01$.
- Percentage of unigrams, bigrams and trigrams in frequency quartiles 1 (lower frequency words) and 4 (higher frequency words) in the source language extracted from the source side of the SMT parallel corpus.
- Percentage of unigrams in the source sentence seen in the source side of the SMT parallel corps.

These features were used to train a Support Vector Regression (SVR) algorithm using a Radial Basis Function (RBF) kernel within the SCIKIT-LEARN toolkit.[13] The γ, ϵ and C parameters were optimised via grid search with 5-fold cross validation on the training set, resulting in γ=0.01, ϵ = 0.0825, C = 20. This baseline system has proved robust across a range of language pairs, MT systems, and text domains for predicting various forms of post-editing effort (2012; 2013; 2014; 2015; 2016a).

Word-level baseline system: For Task 2, the baseline features were extracted with the MARMOT tool (Logacheva et al., 2016). These are 28 features that have been deemed the most informative in previous research on word-level QE. 22 of them were taken from the feature set described in (Luong et al., 2014), and had also been used as a baseline feature set at WMT16:

- Word count in the source and target sentences, and source and target token count ratio. Although these features are sentence-level (i.e. their values will be the same for all words in a sentence), the length of a sentence might influence the probability of a word being wrong.
- Target token, its left and right contexts of 1 word.
- Source word aligned to the target token, its left and right contexts of 1 word. The alignments were given by the SMT system that produced the automatic translations.
- Boolean dictionary features: target token is a stopword, a punctuation mark, a proper noun, or a number.
- Target language model features:
  - The order of the highest order ngram which starts and end with the target token.
  - The order of the highest order ngram which starts and ends with the source token.
  - The part-of-speech (POS) tags of the target and source tokens.
  - Backoff behaviour of the ngrams (t_{i-2}, t_{i-1}, t_i), (t_{i-1}, t_i, t_{i+1}), (t_i, t_{i+1}, t_{i+2}), where t_i is the target token (backoff behaviour is computed as described by (2011)).

In addition to that, 6 new features were included which contain combinations of other features, and which proved useful in (Kreutzer et al., 2015; Martins et al., 2016):

- Target word + left context.
- Target word + right context.
- Target word + aligned source word.
- POS of target word + POS of aligned source word.

[12] https://github.com/ghpaetzold/questplusplus

[13] http://scikit-learn.org/

- Target word + left context + source word.
- Target word + right context + source word.

The baseline system models the task as a sequence prediction problem using the Linear-Chain Conditional Random Fields (CRF) algorithm within the CRFSuite tool (Okazaki, 2007). The model was trained using passive-aggressive optimisation algorithm.

We note that this baseline is different from the one used last year. In Section 4.7 we present results comparing this against last year's baseline.

Phrase-level baseline system: The phrase-level system is identical to the one used in last year's shared task. The phrase-level features were also extracted with MARMOT, but they are different from the word-level features. They are based on the sentence-level features in QUEST++.[14] These are the so-called "black-box" features — features that do not use the internal information from the MT system. The baseline uses the following 72 features:

- Source phrase frequency features:
  - average frequency of ngrams (unigrams, bigrams, trigrams) in different quartiles of frequency (the low and high frequency ngrams) in the source side of the SMT parallel corpus.
  - percentage of distinct source ngrams (unigrams, bigrams, trigrams) seen in the source side of the SMT parallel corpus.
- Translation probability features:
  - average number of translations per source word in the phrase as given by the IBM model 1 extracted using the SMT parallel corpus (with different translation probability thresholds: 0.01, 0.05, 0.1, 0.2, 0.5).
  - average number of translations per source word in the phrase as given by the IBM model 1 extracted using the SMT parallel corpus (with different translation probability thresholds: 0.01, 0.05, 0.1, 0.2, 0.5) weighted by the frequency of each word in the source side of the parallel SMT corpus.
- Punctuation features:
  - difference between numbers of various punctuation marks (periods, commas, colons, semicolons, question and exclamation marks) in the source and the target phrases.
  - difference between numbers of various punctuation marks normalised by the length of the target phrase.
  - percentage of punctuation marks in the target or source phrases.
- Language model features:
  - log probability of the source or target phrases based on models built using the source or target sides of the parallel corpus used to train the SMT system.
  - perplexity of the source and the target phrases using the same models as above.
- Phrase statistics:
  - lengths of the source or target phrases.
  - ratio between the source and target phrase lengths.
  - average length of tokens in source or target phrases.
  - average occurrence of target word within the target phrase.
- Alignment features:
  - number of unaligned target words, using the word alignment provided by the SMT decoder.
  - number of target words aligned to more than one source word.
  - average number of alignments per word in the target phrase.
- Part-of-speech features:
  - percentage of content words in the source or target phrases.
  - percentage of words of a particular part of speech tag (verb, noun, pronoun) in the source or target phrases.
  - ratio of numbers of words of a particular part of speech (verb, noun, pronoun) between the source and target phrases.
  - percentage of numbers and alphanumeric tokens in the source or target phrases.

[14] http://www.quest.dcs.shef.ac.uk/quest_files/features_blackbox

- ratio between the percentage of numbers and alphanumeric tokens in the source and target phrases.

This feature set was designed for sentences. We expect that phrases, being sequences of words of varied length, are similar to sentences and can be treated analogously in QE. On the other hand, unlike sentences, phrases are related to their neighbouring phrases, and in this respect they are similar to words. Therefore, analogously to the baseline word-level system, we treat phrase-level QE as a sequence labelling task, and model it using Conditional Random Fields. The phrase-level baseline system is trained with CRFSuite toolkit using passive-aggressive optimisation algorithm.

4.2 Participants

Table 10 lists all participating teams submitting systems to any of the tasks. Each team was allowed up to two submissions for each task and language pair. In the descriptions below, participation in specific tasks is denoted by a task identifier (T1 = task1, T2 = task 2, T3 = task 3).

CDACM (T2, T3): The submissions from CDACM use a recurrent neural network language model (RNN-LM) architecture for word-level QE as described in (Patel and M, 2016), and explore the word-level predictions for phrase-level QE. CDACM's WMT16 submission was modified to add other RNN variants, such as LSTMs, deep LSTMs and GRUs. Another difference with respect to the WMT16 submission is the addition of the predicted history (only previous prediction) and characters of the word as additional features to the RNN model. This modified architecture predicts the label (OK/BAD) in a slot rather than predicting the word as in the case of standard RNN-LMs. The input to the system is a word sequence, similar to the standard RNN-LM. Bilingual models were also used and performed better than monolingual models. The code for these models is freely available.[15]

DCU (T2): DCU's submission is an ensemble of neural MT systems with different input factors, designed to jointly tackle both the automatic post-editing and word-level QE. Word-level features which have proven effective for QE, such as part-of-speech tags and dependency labels are included as input factors to NMT systems. NMT systems using different input representations are ensembled together in a log-linear model which is tuned for the F_1-mult metric using MERT (Och, 2003). The output of the ensemble is a pseudo-reference that is then TER aligned with the original MT to obtain OK/BAD tags for each word in the MT hypothesis.

DFKI (T1): These submissions investigate alternative machine learning models for the prediction of the HTER score on the sentence-level task. Instead of directly predicting the HTER score, the systems use a single-layer perceptron with four outputs that jointly predict the number of each of the four distinct post-editing operations that are then used to calculate the HTER score. This also gives the possibility to correct invalid (e.g. negative) predicted values prior to the calculation of the HTER score. The two submissions use the baseline features and the English-German submission also uses features from (Avramidis, 2017a).

JXNU (T1): The JXNU submissions use features extracted from a neural network, including embedding features and cross-entropy features of the source sentences and their machine translations. The sentence embedding features are extracted through global average pooling from word embedding, which are trained using the WORD2VEC toolkit. The sentence cross-entropy features are calculated by a recurrent neural network language model. They experimented with different sentence embedding dimensions of the source sentences and translation outputs, as well as different sizes of the training corpus. The experimental results show that the neural network features lead to significant improvements over the baseline, and that combining the neural network features with baseline features leads to further improvement.

POSTECH (T1, T2, T3): POSTECH's submissions to the sentence/word/phrase-level QE tasks are based on predictor-estimator architecture (Kim et al., 2017; Kim and Lee, 2016), which is the two-stage end-to-end

[15] https://github.com/patelrajnath/rnn4nlp

ID	Participating team
CDACM	Centre for Development of Advanced Computing, India (Patel and M, 2016)
DCU	Dublin City University (Hokamp, 2017)
DFKI	German Research Centre for Artificial Intelligence, Germany (Avramidis, 2017b)
JXNU	Jiangxi Normal University, China (Chen et al., 2017)
POSTECH	Pohang University of Science and Technology, Republic of Korea (Kim et al., 2017)
RTM	Referential Translation Machines, Turkey (Biçici, 2017)
SHEF	University of Sheffield, UK (Blain et al., 2017; Paetzold and Specia, 2017)
UHH	University of Hamburg, Germany (Duma and Menzel, 2017)
Unbabel	Unbabel, Portugal (Martins et al., 2017b)

Table 10: Participants in the WMT17 quality estimation shared task.

neural QE model. The predictor-estimator architecture consists of two types of stacked neural network models: 1) a word prediction model based on bidirectional and bilingual recurrent neural network language model trained on additional large-scale parallel corpora and 2) a neural quality estimation model trained on quality-annotated noisy parallel corpora. To jointly learn the two-stage model, a stack propagation method was applied (Zhang and Weiss, 2016). In addition, a "multilevel model" was developed where a task-specific predictor-estimator model was trained using not only task-specific training examples but also all the other training examples of QE subtasks. All the submitted runs are ensembles that combine a set of neural models, trained under different settings of varying dimensionalities and shuffling of training examples.

RTM (T1, T2, T3): The RTM systems are improved versions over WMT16's RTM submissions which average prediction scores from different models using weights based on their training performance to improve the overall test performance. They also use new features representing substring distances, punctuation tokens, character n-grams, and alignment crossings.

SHEF (T1, T2, T3): The SHEF team participated in all the three sub-tasks. For task 1, two types of systems were submitted: CNN and QUEST-EMB. The CNN submissions are based on convolutional neural networks. The system first transforms the source and target sentences into sequences of character embeddings, and then passes them through a series of deep parallel stacked convolution/max pooling layers. The baseline features are provided through a multi-layer perceptron, and then concatenated with the character-level information. Finally, the concatenation is passed onto another multi-layer perceptron and the very last layer outputs HTER values. The two submissions differ in the the use of standard (CNN+BASE-Single) and multi-task learning (CNN+BASE-Multi) for training. The QUEST-EMB submission follows the word embeddings approach used by (Scarton et al., 2016) for document-level QE. Here in-domain word embeddings are used instead of embeddings obtained general purpose data (same as in task 2, below). Word embeddings were averaged to generate a single vector for each sentence. Source and target word embeddings were then concatenated with the baseline features and given to an SVM regressor for model building.

For the word-level task SHEF investigated a new approach based on predicting the strength of the lexical relationships between the source and target sentences (BMAPS). Following the work by (Madhyastha et al., 2014), a bilinear model is trained from three matrices corresponding to the training data, the development set and a "truth" matrix between them, which is built from the word alignments and the gold labels to indicate which lexical items form a pair, and whether or not their lexical relation is OK or BAD. The first two matrices are built from 300 dimension word vectors computed with pretrained in-domain word embeddings. They train their model over 100 iterations with the l_2 norm as regulariser and using the *forward-backward splitting* algorithm (FOBOS) (Duchi and Singer, 2009) as optimisation method. They report results considering the word and its context versus the word in isolation, as well as variants with and without the gold labels at training time.

Finally, for the phrase-level task, SHEF made use of predictions generated by BMAPS for task 2 and the phrase labelling approaches in (Blain et al., 2016). These approaches use the number of BAD word-level predictions in a phrase: an optimistic version labels the phrase as OK if at least half of the words in it are predicted to be OK, and a super-pessimistic version labels the phrase as BAD if any word is in is predicted to be BAD.

UHH (T1): The UHH-STK submission is based on sequence and tree kernels applied on the source and target input data for predicting the HTER score. The kernels use a back-translation of the MT output into the source language as an additional input data representation. Further hand-crafted features were defined in the form of the scores of the kernel functions applied on the pair of source and back-translation sentences. The submitted runs outperformed the baseline systems for both language pairs.

Unbabel (T1, T2): For word level, the "stacked" system stacks a linear and a neural model similar to the ones submitted by Unbabel at WMT16. The "full-stacked-src-mt" system incorporates the output of an APE system, converted to OK/BAD tags, as an additional feature, similar to their work in (Martins et al., 2017a). The sentence-level submissions use and normalise the word-level predictions as percentage of words edited to generate an HTER score.

4.3 Datasets

One of the main differences between this year's and previous years' tasks is the considerably larger size of human-labelled datasets made available to participants for training. Whereas the last year we released a corpus of $12,000$ instances (plus $1,000$ and $2,000$ for development and test, respectively), this year this figure was doubled. In contrast to last year, we also provide datasets for two language pairs.

The structure used for the data have been the same since WMT15. Each data instance consists of (i) a source sentence, (ii) its automatic translation into the target language, (iii) the manually post-edited version of the automatic translation, (iv) a free reference translation of the source sentence. Post-edits are used to extract labels for the different levels of granularity, which allows using the same datasets for all three QE tasks.

The first dataset contains texts in **IT domain** translated **from English into German**. This is a superset of the last year's data: $11,000$ sentences from the same source were added to the training set. Their translations were produced using the same statistical MT system and post-edited by professional translators who are native speakers of German. The dataset statistics are outlined in Table 11.

The second dataset belongs to **pharmaceutical domain** and provides translations **from German into English**. It contains $25,000$ instances for training. Analogously to the IT dataset, automatic translations were generated with a statistical MT system and post-edited by professional translators. The dataset statistics are shown in Table 12. The Table shows another feature of this dataset: it contains much fewer errors than the IT one.

	Sentences	Words	% of BAD words
Training	23,000	404,198	20.55
Development	1,000	19,487	19.55
Test	2,000	35,577	19.70

Table 11: Statistics of the English–German dataset.

	Sentences	Words	% of BAD words
Training	25,000	453,666	12.55
Development	1,000	18,152	11.71
Test	2,000	36,119	11.52

Table 12: Statistics of the German–English dataset.

4.4 Task 1: Predicting sentence-level quality

This task consists in scoring (and ranking) translation sentences according to the proportion of their words that need to be fixed. HTER (Snover et al., 2006b) is used as quality score, i.e. the minimum edit distance between the machine translation and its manually post-edited version.

Labels HTER labels were computed using the TERCOM tool[16] with default settings (tokenised, case insensitive, exact matching only), with scores capped to 1.

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

Evaluation Evaluation was performed against the true HTER label and/or ranking, using the following metrics:

- Scoring: Pearson's r correlation score (primary metric, official score for ranking submissions), Mean Average Error (MAE) and Root Mean Squared Error (RMSE).

- Ranking: Spearman's ρ rank correlation and DeltaAvg.

Statistical significance on Pearson r was computed using the William's test.[17]

Results Tables 13 and 14 summarise the results for Task 1 on German–English and English–German datasets, respectively, ranking participating systems best to worst using Pearson's r correlation as primary key. Spearman's ρ correlation scores should be used to rank systems for the ranking variant.

The top three systems are the same for both datasets, and the ranking of systems according to their performance is similar for both datasets. They are all based on neural models that first model the problem of word-level prediction and then somehow generalise such predictions for sentence level QE, either by using them directly (Unbabel) or building a model from word to sentence-level prediction (POSTECH). We also note that the majority of the systems perform better than the baseline, although five submissions are not significantly different from it.

4.5 Task 2: Predicting word-level quality

This task evaluates the extent to which we can detect word-level errors in MT output. Often, the overall quality of a translated segment is significantly harmed by specific errors in a small proportion of the words. Various classes of errors can be found in translations, but for this task we consider all error types together, aiming at making a binary distinction between correct (OK) and incorrect (BAD) tokens.

Labels The binary labels for the datasets (OK and BAD) were derived automatically from the TERCOM tool with default settings and disabled shifts (option "-d 0"). We aligned automatically translated sentences with their post-edited version and labelled each word in the automatic translation with an edit operation: insertion, deletion, substitution or no edit (correct word). We mark each edited word as BAD, and the remainingn as OK.

Evaluation Analogously to the last year's task, the primary evaluation metric is the multiplication of F_1-scores for the OK and BAD classes, denoted as F_1-mult. Unlike previously used F_1-BAD score this metric is not biased towards "pessimistic" labellings. We also report F_1-scores for individual classes for completeness. We test the significance of the results using randomisation tests (Yeh, 2000) with Bonferroni correction (Abdi, 2007).

Results The results for Task 2 are summarised in Tables 15 and 16, ordered by the F_1-mult metric.

The top two systems are the same as for the sentence-level task. This is perhaps not surprising since these are essentially word-level predictors: POSTECH and Unbabel. These along with DCU's submissions (which were specifically designed for the English–German word-level task), are all based on neural models.

4.5.1 Word-level predictions for sentence-level QE

Given that some submissions to the sentence-level task which were actually based on word-level predictions performed very well at sentence level, here we study the performance of *all* teams participating in the word-level task for sentence-level prediction. The percentage of words labelled as BAD in a sentence can essentially be seen as a sentence-level HTER score. Participants were also invited to submit an additional word-level system tuned to optimise sentence-level scores, but we are not aware of systems that did so.

In order to obtain sentence-level scores from word-level predictions we computed HTER for each sentence in the test set as the percentage of words classified as BAD. We then evaluated the submissions in terms of sentence-level metrics: Pearson correlation, MAE, RMSE. Table 17 shows the performance of the word-level systems on the sentence-level task for the German–English dataset and their comparison with the participants of the Task 1. It can be clearly seen that word-level predictions are very close to sentence-level ones: systems of different levels are well distributed along the ranked list.

The submissions by POSTECH and Unbabel show that word-level and sentence-level systems

[17] `https://github.com/ygraham/mt-qe-eval`

Model	**Pearson r**	MAE	RMSE	Spearman ρ	DeltaAvg
● POSTECH/Combined-MultiLevel-Ensemble	0.728	0.091	0.133	0.691	10.64
POSTECH/SingleLevel-Ensemble	0.715	0.094	0.136	0.669	10.44
Unbabel/full-stacked-src-mt	0.626	0.121	0.179	0.613	9.74
RTM/RTM-MIX	0.600	0.109	0.157	0.570	8.94
RTM/RTM-TREE	0.585	0.119	0.158	0.573	9.18
Unbabel/stacked	0.580	0.106	0.170	0.574	7.72
SHEF/QUEST-EMB-SCALE	0.558	0.121	0.161	0.561	8.79
JXNU/Emb+RNNLM+QuEst+SVM	0.531	0.130	0.167	0.520	8.62
UHH/STK1	0.503	0.137	0.172	0.503	8.17
UHH/STK2	0.489	0.140	0.175	0.482	7.97
BASELINE	0.441	0.128	0.175	0.446	6.81
DFKI/SLP4	0.398	0.123	0.188	0.396	5.82
SHEF/CNN+BASE-Single	0.390	0.136	0.179	0.388	6.39
SHEF/CNN+BASE-Multi	0.350	0.162	0.202	0.387	6.41

Table 13: Official results of the WMT17 Quality Estimation Task 1 for the German–English dataset. The winning submission is indicated by a ● and is statistically significantly different from all others. Submissions in the grey area are those which are not significantly different from the baseline.

Model	**Pearson r**	MAE	RMSE	Spearman ρ	DeltaAvg
● POSTECH/Combined-MultiLevel-Ensemble	0.695	0.102	0.137	0.725	12.32
POSTECH/SingleLevel-Ensemble	0.673	0.107	0.141	0.703	11.98
Unbabel/full-stacked-src-mt	0.641	0.128	0.169	0.652	11.36
Unbabel/stacked	0.589	0.129	0.176	0.610	10.28
JXNU/Emb+RNNLM+QuEst+SVM	0.522	0.126	0.163	0.545	9.54
UHH/STK2	0.509	0.130	0.166	0.534	9.41
UHH/STK1	0.508	0.129	0.165	0.533	9.49
SHEF/QUEST-EMB-SCALE	0.496	0.126	0.166	0.513	8.96
RTM-MIX	0.454	0.130	0.171	0.477	8.64
RTM-PLS-GBR	0.430	0.131	0.173	0.452	8.23
SHEF/CNN+BASE-Single	0.416	0.135	0.174	0.444	8.13
SHEF/CNN+BASE-Multi	0.402	0.135	0.178	0.452	8.16
BASELINE	0.397	0.136	0.175	0.425	7.45
DFKI/SLP4	0.113	0.153	0.204	0.136	2.5

Table 14: Official results of the WMT17 Quality Estimation Task 1 for the English–German dataset. The winning submission is indicated by a ● and is statistically significantly different from all others. Submissions in the grey area are those which are not significantly different from the baseline.

Model	**F_1-mult**	F_1-BAD	F_1-OK
● POSTECH/Combined-MultiLevel-Ensemble	0.535	0.569	0.940
● Unbabel/full-stacked-src	0.529	0.562	0.941
POSTECH/SingleLevel-Ensemble	0.516	0.552	0.936
Unbabel/stacked	0.466	0.497	0.936
BASELINE	0.342	0.365	0.939
CDACM/RNN	0.333	0.370	0.900
RTM/s4-RTM-GLMd	0.329	0.350	0.939
SHEF/BMAPS-unigram	0.088	0.210	0.419
SHEF/BMAPS-nolabel-unigram	0.082	0.209	0.391

Table 15: Official results of the WMT17 Quality Estimation Task 2 for the German–English dataset. The winning submissions are indicated by a ● and are statistically significantly different from all others. Submissions in the grey area are those which are not significantly different from the baseline.

Model	**F_1-mult**	F_1-BAD	F_1-OK
• POSTECH/Combined-MultiLevel-Ensemble	0.568	0.628	0.904
• Unbabel/full-stacked-src-mt	0.566	0.625	0.906
• DCU/SRC-APE-QE-TUNED	0.559	0.614	0.910
• DCU/AVG-ALL	0.556	0.611	0.910
POSTECH/SingleLevel-Ensemble	0.543	0.607	0.894
Unbabel/stacked	0.512	0.581	0.882
CDACM/RNN	0.370	0.457	0.809
BASELINE	0.361	0.407	0.886
RTM/s5-RTM-GLMd	0.285	0.322	0.884
RTM/s4-RTM-GLMd	0.261	0.293	0.889
SHEF/BMAPS-unigram	0.097	0.302	0.322
SHEF/BMAPS-nolabel-unigram	0.157	0.325	0.484

Table 16: Official results of the WMT17 Quality Estimation Task 2 for the English–German dataset. The winning submissions are indicated by a • and are statistically significantly different from all others. Submissions in the grey area are those which are not significantly different from the baseline.

trained on the same data using the same (or similar) methods yield very close results: the POSTECH sentence-level systems occupy the first two positions in the list, while their word-level systems follow. The corresponding word-level and sentence-level systems by Unbabel are even closer, their differences are not statistically significant. This is expected since Unbabel's submission to the sentence-level task was based on their predictions for the word-level task. Finally, the baselines for the two task do not show significant differences in their performance either, although they are based on very different features and models.

Overall, these results suggest that word-level QE models can indeed be successfully used to predict sentence-level quality of translation. Additionally, sentence-level metrics proved suitable for the evaluation of word-level QE models (the rankings of word-level submissions produced by F_1-mult and Pearson r metrics have correlation coefficient of 0.96). Results for the English–German task show the same trend.

4.6 Task 3: Predicting phrase-level quality

This level of granularity was first introduced in the shared task at WMT16. The goal is to predict MT quality at the level of phrases.

Labels The phrase-level QE task requires segmenting training and test sentences into phrases. We used the segmentation produced by the SMT system which generated automatic translations for the datasets. The phrase-level labels were produced from binary word-level labels: we labelled a phrase as OK if all words in it were correct (OK words). Any phrase with one or more BAD words was labelled as BAD.

Evaluation In contrast to the last year's phrase-level shared task, where we used word-level metrics to evaluate phrase-level submissions, this time we resort to phrase-level F_1 scores. The reason for that is that the word-level metrics were unable to differentiate between various systems. Therefore, here our primary metric is the phrase-level version of F_1-mult, and we also report phrase-level F_1-BAD and F_1-OK. Statistical significance was computed using randomised test with Bonferroni correction as in task 2.

Results The results of the phrase-level task are represented in Tables 18 and 19. These results follow from those for the word-level task, with POSTECH showing significantly better results overall.

4.7 Discussion

In what follows, we discuss the main findings of this year's shared task based on the goals we had previously identified for it.

Larger training data To test the effectiveness of larger (domain-specific and professionally annotated) datasets, we increase the size of last year's training set for English–German. In order to check if the increased training data size helps improve the systems' performance we compare the baseline systems for all tasks trained on last year's versus this year's dataset, with parameters optimised on the same development sets.

Model	**Pearson** r	MAE	RMSE
● POSTECH/Combined-MultiLevel-Ensemble	0.728	0.091	0.133
POSTECH/SingleLevel-Ensemble	0.715	0.094	0.136
word POSTECH/Combined-MultiLevel-Ensemble	0.687	0.092	0.149
word POSTECH/SingleLevel-Ensemble	0.674	0.095	0.153
Unbabel/full-stacked-src-mt	0.626	0.121	0.179
word Unbabel/full-stacked-src-mt	0.625	0.147	0.242
RTM/RTM-MIX	0.600	0.109	0.157
RTM/RTM-TREE	0.585	0.119	0.158
Unbabel/stacked	0.580	0.106	0.170
word Unbabel/stacked	0.580	0.147	0.242
SHEF2/QUEST-EMB-SCALE	0.558	0.121	0.161
JXNU/Emb+RNNLM+QuEst+SVM	0.531	0.130	0.167
UHH/STK1	0.503	0.137	0.172
UHH/STK2	0.489	0.140	0.175
word BASELINE	0.455	0.118	0.197
word CDACM/RNN	0.450	0.132	0.198
BASELINE	0.441	0.128	0.175
word RTM/s4-RTM-GLMd	0.425	0.122	0.201
DFKI/SLP4	0.398	0.123	0.188
SHEF1/CNN+BASE-Single	0.390	0.136	0.179
SHEF1/CNN+BASE-Multi	0.350	0.162	0.202
word SHEF/BMAPS-nolabel-unigram	0.180	0.592	0.628
word SHEF/BMAPS-unigram	0.167	0.574	0.613

Table 17: Additional results of the WMT17 Quality Estimation Task 1 for the German–English dataset: using for the word-level predictions for sentence-level QE, evaluated for scoring. The winning submission is indicated by a ● and is statistically significantly different from all others. Submissions in the grey area are those which are not significantly different from the baselines. The word-level systems are denoted with prefix **word**.

Model	$\mathbf{F_1}$**-mult**	F_1-BAD	F_1-OK
● POSTECH/PredictorEstimator-Combined-MultiLevel-Ensemble	0.561	0.615	0.912
POSTECH/PredictorEstimator-SingleLevel-Ensemble	0.543	0.599	0.906
CDACM/RNN	0.381	0.444	0.858
BASELINE	0.360	0.397	0.907
RTM/s5-RTM-GLMd	0.284	0.312	0.908
RTM/s4-RTM-GLMd	0.278	0.306	0.908
SHEF/BMAPS-unigram-opti	0.141	0.299	0.473
SHEF/BMAPS-unigram-nolabel-opti	0.132	0.300	0.440

Table 18: Official results for the WMT17 Quality Estimation Task 3 for the German-English data. The winning submission is indicated by a ● and is statistically significantly different from all others. The gray area indicates the submissions whose results are not statistically different from the baseline.

Model	$\mathbf{F_1}$**-mult**	F_1-BAD	F_1-OK
● POSTECH/PredictorEstimator-Combined-MultiLevel-Ensemble	0.586	0.679	0.863
POSTECH/PredictorEstimator-SingleLevel-Ensemble	0.549	0.652	0.843
CDACM/RNN	0.391	0.535	0.731
BASELINE	0.327	0.402	0.814
SHEF/BMAPS-unigram-opti	0.226	0.409	0.553
SHEF/BMAPS-unigram-nolabel-opti	0.148	0.388	0.380

Table 19: Official results for the WMT17 Quality Estimation Task 3 for the English–German data. The winning submission is indicated by a ● and is statistically significantly different from all others. The gray area indicates the submissions whose results are not statistically different from the baseline.

In Table 20 we show the performance of the baseline systems for all tasks trained on the WMT16 and WMT17 English–German datasets and tested on the WMT16 test set. The performance improves for all tasks when using the WMT17 training set, which is much larger. However, the gain for the word-level and phrase-level tasks is smaller than that for sentence level. For the word-level task, we also include experiments with the WMT16 baseline system, which was simpler than the WMT17 baseline system. We observe larger improvement from the new word-level features which we included in this year's baseline system than from the larger training set. This suggests that better features/models can lead to larger performance gains than more data, at least for the word-level task.

2016 word-level baseline			
Training set	F_1-mult	F_1-BAD	F_1-OK
2016 data	0.324	0.368	0.880
2017 data	0.335	0.378	0.886
2017 word-level baseline			
Training set	F_1-mult	F_1-BAD	F_1-OK
2016 data	0.341	0.384	0.887
2017 data	0.360	0.404	0.892
Phrase-level baseline			
Training set	F_1-mult	F_1-BAD	F_1-OK
2016 data	0.311	0.389	0.799
2017 data	0.328	0.403	0.812
Sentence-level baseline			
Training set	Pearson r	MAE	RMSE
2016 data	0.351	0.135	0.184
2017 data	0.397	0.136	0.175

Table 20: Comparison of baseline English–German systems trained on WMT16 and WMT17 datasets (tested on the WMT16 test set) for all tasks.

Progress over years Progress over years is a difficult factor to measure. We attempted to do so this year for the first time given the similarity between the tasks this and last year for the English–German data. We do so by requesting participants in this year's task to submit results using their WMT17 systems on the WMT16 test sets. We note however that this comparison is also affected by the increased size of the training set for this language pair in the current edition of the task. Therefore, the WMT17 systems may be better systems because of better techniques but also because of larger amounts of training data.

In Table 21 we compare the results from WMT16 and WMT17 systems on the WMT16 test set at sentence level, where WMT16 systems are highlighted in cyan background. Overall, it can be clearly seen that WMT17 systems perform better: last year's top system is only the 4th best compared to the WMT17 submissions, and half of WMT16 participants are below this year's baseline. It is important to note that the baseline performs much better than last year because of the additional training data – as shown in Table 20 – since the baseline system itself did not change.

Table 22 shows the results for word-level systems, which indicates a similar trend: systems also improved from last year's submissions, with last year's winner being outperformed by four other systems, and the majority of WMT16 participants performing closely to this year's baseline (which we note is a stronger model than last year's baseline as previously discussed).

Finally, the same trend is observed when comparing phrase-level systems submitted to WMT16 and WMT17 in Table 23. The only difference is that although the new data improved the performance of the phrase-level baseline system, this improvement did not change its position in the systems ranking.

Overall, the (Person r and F_1-mult) scores of the winning submissions this year is much higher than in last year's results, which we believe to be a combination of better techniques as well as better (larger) data.

The progress of state-of-the-art QE models can also be tracked by the performance of recurring participants: the results of systems by POSTECH (tasks 1, 2, 3) and CDACM (task 2) teams are better this year.

We note the increasing popularity of neural networks and their improving performance for QE: although some of the last year's winners (e.g. YSDA team which won the sentence-level task) did not use neural networks, all WMT17 winners and the majority of best-performing systems use neural networks for model building.

Languages and domains To study the effect of language direction and domain, we provided two datasets created in similar ways, but for different domains and language directions, as was previously mentioned. The QE performance on these datasets varies considerably, with German–English showing higher scores for the sentence-

Model	**Pearson** r	MAE	RMSE
• POSTECH/PredictorEstimator-Combined-MultiLevel-Ensemble	0.714	0.096	0.134
POSTECH/PredictorEstimator-SingleLevel-Ensemble	0.686	0.101	0.139
JXNU/Emb+RNNLM+QuEst+SVM	0.527	0.122	0.163
• YSDA/SNTX+BLEU+SVM	0.525	12.30	16.41
UHH/STK2	0.524	0.124	0.162
UHH/STK1	0.516	0.123	0.163
SHEF/QUEST-EMB-SCALE	0.499	0.124	0.167
POSTECH/SENT-RNN-QV2	0.460	13.58	18.60
SHEF-LIUM/SVM-NN-emb-QuEst	0.451	12.88	17.03
POSTECH/SENT-RNN-QV3	0.447	13.52	18.38
SHEF-LIUM/SVM-NN-both-emb	0.430	12.97	17.33
SHEF/CNN+BASE-Single	0.421	0.131	0.174
UGENT-LT3/SCATE-SVM2	0.412	19.57	24.11
BASELINE 2017	0.399	0.132	0.175
SHEF/CNN+BASE-Multi	0.397	0.135	0.184
UFAL/MULTIVEC	0.377	13.60	17.64
RTM/RTM-FS-SVR	0.376	13.46	17.81
UU/UU-SVM	0.370	13.43	18.15
UGENT-LT3/SCATE-SVM1	0.363	20.01	24.63
RTM/RTM-SVR	0.358	13.59	18.06
BASELINE 2016	0.351	13.53	18.39
SHEF/SimpleNets-SRC	0.320	13.92	18.23
SHEF/SimpleNets-TGT	0.283	14.35	18.22
RTM-PLS-GBR	0.163	0.150	0.192
RTM-TREE	0.155	0.148	0.190
DFKI/SLP4	0.132	0.154	0.206

Table 21: Comparison of official results of WMT17 and WMT16 sentence-level QE task on the English–German WMT16 test set. The winning submission is indicated by a • and is statistically significantly different from all others. WMT16 systems are highlighted with cyan.

level task, both in terms of the baseline systems the winning submissions, and English–German showing generally higher scores for the word and phrase-level tasks (except for the baseline system in the phrase-level task). Even though the performance scores may not be directly comparable, we can make some interesting observations. We believe that the main reasons for these differences are related to the general quality of the MT systems and – as a consequence – the distribution of quality labels in the QE datasets, and – to a lesser extent – the sizes of the QE training sets, which are slightly different (see Tables 11 and 12).

The quality of the translations in each dataset is very different. As shown in Tables 11 and 12, the German–English dataset contains much fewer errors. Indeed, when building the SMT systems that generated these translations, we observed very different BLEU scores: 35.9 for English–German (IT domain), and 53.4 for German–English (Pharma domain). This difference in quality is not due to training settings, since these were the same for both datasets, except that for English–German the SMT training set was much larger (7.2 vs 2.09 million sentences). Details on the SMT models and data used to build such models are given in (Specia et al., 2017a). In addition to the well-known fact that translating into English normally leads to better quality than translating from English, we hypothesise that this difference could be due to higher token repetition rate in the German–English dataset. The difference in quality was confirmed by the average HTER score obtaining from the post-editing of these test sets: 0.25 for English–German and 0.19 for German–English.[18]. The fact that the German–English dataset contains fewer errors makes it harder for the word and phrase-level tasks to achieve high F_1-mult as

[18] We note that these BLEU and HTER scores were measured on a superset of this data, as described in (Specia et al., 2017a)

Model	**F$_1$-mult**	F$_1$-BAD	F$_1$-OK
● POSTECH/Combined-MultiLevel-Ensemble	0.581	0.637	0.913
● DCU/SRC-APE-QE-TUNED	0.575	0.627	0.917
● DCU/AVG-ALL	0.573	0.625	0.917
POSTECH/SingleLevel-Ensemble	0.561	0.619	0.906
● Unbabel/ensemble	0.495	0.560	0.885
Unbabel/linear	0.463	0.529	0.875
UGENT-LT3/SCATE-RF	0.411	0.492	0.836
CDACM/RNN	0.391	0.469	0.833
UGENT-LT3/SCATE-ENS	0.381	0.464	0.821
POSTECH/WORD-RNN-QV3	0.380	0.447	0.850
POSTECH/WORD-RNN-QV2	0.376	0.454	0.828
UAlacant/SBI-Online-baseline	0.367	0.456	0.805
BASELINE 2017	0.360	0.404	0.892
CDACM/RNN	0.353	0.419	0.842
SHEF/SHEF-MIME-1	0.338	0.403	0.839
SHEF/SHEF-MIME-0.3	0.330	0.391	0.845
BASELINE 2016	0.324	0.368	0.880
RTM/s5-RTM-GLMd	0.308	0.349	0.882
RTM/s5-RTM-GLMd	0.305	0.353	0.865
UAlacant/SBI-Online	0.290	0.406	0.715
RTM/s4-RTM-GLMd	0.286	0.326	0.878
RTM/s4-RTM-GLMd	0.273	0.307	0.888
SHEF/BMAPS-unigram	0.158	0.316	0.501
SHEF/SHEF/BMAPS-nolabel-unigram	0.098	0.296	0.330

Table 22: Comparison of official results of WMT17 and WMT16 word-level QE task on the English–German WMT16 test set. Winning submissions are indicated by a ● and are statistically significantly different from all others. WMT16 systems are highlighted with cyan.

Model	**F$_1$-mult**	F$_1$-BAD	F$_1$-OK
● POSTECH/Combined-MultiLevel-Ensemble	0.603	0.693	0.869
POSTECH/SingleLevel-Ensemble	0.562	0.662	0.849
CDACM/RNN	0.403	0.541	0.744
POSTECH/RNN-QV3	0.393	0.518	0.759
POSTECH/RNN-QV2	0.388	0.504	0.771
CDACM/RNN	0.378	0.500	0.756
USFD2/CONTEXT	0.364	0.467	0.780
USFD2/W&SLP4PT	0.363	0.475	0.764
RTM/s5-RTM-GLMd	0.342	0.420	0.814
RTM/s4-RTM-GLMd	0.336	0.411	0.817
RTM/s5-RTM-GLMd	0.331	0.413	0.802
BASELINE 2017	0.328	0.403	0.812
BASELINE 2016	0.311	0.389	0.799
RTM/s4-RTM-GLMd	0.306	0.376	0.815
UAlacant/SBI-Online-baseline	0.275	0.502	0.547
SHEF/BMAPS-unigram-opti	0.233	0.415	0.562
SHEF/BMAPS-unigram-nolabel-opti	0.149	0.398	0.373
UAlacant/SBI-Online	0.146	0.456	0.320

Table 23: Comparison of official results of WMT17 and WMT16 phrase-level QE task on the English–German WMT16 test set. The winning submission is indicated by a ● and is statistically significantly different from all others. WMT16 systems are highlighted with cyan.

the models will have a strong bias towards predicting words or phrases as OK. In fact, if we take the word-level task, the difference between F_1-BAD and F_1-OK scores is much more noticeable for German–English (0.569 vs 0.940, respectively – Table 15) than for English–German (0.628 vs 0.904, respectively – Table 16), showing that the systems tend to overpredict OK labels for German–English. The same applies to the phrase-level task. For the sentence-level task, the skewed distribution towards good quality translations does not have the same effect, perhaps due to the prediction of an aggregated (HTER) score and the metric used for evaluation.

Additional evidence To investigate the utility of detailed information logged during post-editing, we offered to participants other sources of information: post-editing time, keystrokes, and actual edits. Surprisingly, no participating system requested these additional labels. The DFKI submission re-created some of this information by further annotating words with the actual edit operations, as obtained from the HTER alignments. Instead of predicting the HTER score, the systems attempted to predict the number of each of the four post-editing operations (add, replace, shift, delete) at the sentence level. However, this did not lead to positive results. In future editions of the task, we plan to make this detailed post-editing information available again and suggest clear ways of using it.

5 Automatic Post-editing Task

The WMT shared task on MT automatic post-editing (APE), this year at its third round at WMT, aims to evaluate systems for the automatic correction of errors in a machine translated text. As pointed out by (Chatterjee et al., 2015b), from the application point of view the task is motivated by its possible uses to:

- Improve MT output by exploiting information unavailable to the decoder, or by performing deeper text analysis that is too expensive at the decoding stage;
- Cope with systematic errors of an MT system whose decoding process is not accessible;
- Provide professional translators with improved MT output quality to reduce (human) post-editing effort;
- Adapt the output of a general-purpose MT system to the lexicon/style requested in a specific application domain.

The third round of the APE task proposed to participants the same general evaluation framework of the previous ones (Bojar et al., 2015, 2016a). It consists in a “black box” scenario in which the MT system that produced the translations is unknown to the participants and cannot be modified.

This year the task has been extended by including German-English as a new language direction in addition to English-German, which was the only language pair covered in the 2016 round. For both directions, participants operated with domain-specific data (`information technology` for EN-DE and `pharmacological` for DE-EN),[19] with post-edits collected from professional translators.[20] All data has been provided by the European Project QT21.[21]

As in 2016, TER and BLEU computed between automatic and human post-edits have been respectively used as primary and secondary evaluation metrics. In continuity with the previous round, a manual evaluation has also been carried out to gain further insights on final output quality. However, while in 2016 Appraise[22] (Federmann, 2012) was employed for manual evaluation, this year the German to English evaluation was carried out via direct human assessment (Graham et al., 2016) and quality controlled crowd-sourcing on Amazon's Mechanical Turk[23], while the English to German evaluation was completed, again via direct assessment, but translation students were employed as opposed to crowd-sourcing.

In terms of participants and submitted runs, this year's round replicated the success of the 2016 edition. On English-German we had 7 participants (one more than in 2016), with a total of 15 submitted runs. On German-English (a more challenging direction due to a much higher quality of the original MT output), we had 2 participants, with a total of 5 submitted runs.

Building on the recent success of neural ap-

[19] As opposed to the general `news` domain data used in the first round, which proved to be more difficult to handle due to scarce repetitiveness.

[20] As opposed to the less coherent crowdsourced material used in the first round.

[21] `http://www.qt21.eu/`

[22] `https://github.com/cfedermann/Appraise`

[23] `https://www.mturk.com`

proaches to APE, this year all the submissions relied on neural end-to-end solutions. The adoption of multi-source models (able to combine information from raw MT output and the original source text) and the extensive use of available synthetic data (to increase the size of the training set) are other traits common to several systems.

On both directions, all participants managed to beat the baseline, at least with their primary submission. Top results achieved impressive improvements up to -4.9 TER and +7.6 BLEU points on English-German and smaller, but statistically significant gains up to -0.25 TER and +0.3 BLEU on German-English. The manual evaluation of participants' primary submissions confirmed the jump in performance of this year's systems in the English-German task. Although all of them are still below human quality, the gap has been reduced with respect to the 2016 round, three systems are almost on par in the top tier (last year it was only one) and the improvements over the baseline are significantly better than the original MT output prior to post-editing for all participants (last year this was only true for the top submission).

5.1 Task description

Similar to previous years, participants were provided with training and development data consisting of (*source*, *target*, *human post-edit*) triplets, and were asked to return automatic post-edits for a test set of unseen (*source*, *target*) pairs.

5.1.1 Data

Previous rounds of the APE task suggested (Bojar et al., 2015) and confirmed (Bojar et al., 2016a) the dependence of system results on data repetitiveness. In the 2015 pilot task, dealing with "general-domain" `news` data and crowdsourced post-edits proved to be very difficult due to data sparsity issues that prevented participants to learn from the training set useful correction patterns reapplicable to the test set. In 2016, the switch to more repetitive (in other terms, less sparse) domain-specific data post-edited by professional translators resulted in a higher applicability of the learned correction patterns. The effect of this switch was made evident by final results: while none of the submitted runs was able to beat the baseline in the pilot round, more than half of the submissions significantly outperformed it in 2016. Based on these outcomes, and to give stability to a relatively young task, also this year we opted for the adoption of domain-specific data post-edited by professionals for both language directions.

Training and development sets consist of (*source*, *target*, *human post-edit*) triplets in which:

- The source (SRC) is a tokenized sentence with length between 3 and 30 tokens;
- The target (TGT) is a tokenized translation of the source. Translations were obtained from statistical MT systems.[24] This information, however, was unknown to participants, for which the MT system was a black-box.
- The human post-edit (PE) is a manually-revised version of the target, done by professional translators.

Test data consists of (*source*, *target*) pairs having similar characteristics of those in the training set. Human post-edits of the test target instances were left apart to measure system performance.

English-German data were drawn from the `Information Technology` (IT) domain. Training and test sets respectively contain 11,000 and 2,000 triplets. The data released for the 2016 round of the task (15,000 instances) and the artificially generated post-editing triplets (4 million instances) used by last year's winning system (Junczys-Dowmunt and Grundkiewicz, 2016) were also provided as additional training material.

German-English data were drawn from the `Pharmacological` domain. Training and development sets respectively contain 25,000 and 1,000 triplets, while the test set consists of 2,000 instances.

Table 24 provides some basic statistics about the data (the same used for the sentence-level quality estimation task), which has been released by the European Project QT21 (Specia et al., 2017b).[25] In addition, Tables 25 and 26 provide a view of the data from a task difficulty standpoint. Table 25 shows the repetition rate (RR) values of the data sets released in the three rounds

[24] We used phrase-based MT systems trained with generic and in-domain parallel training data, leveraging prereordering techniques (Herrmann et al., 2013), and taking advantage of POS and word class-based language models.

[25] For both language directions, the source sentences and reference translations were provided by TAUS (https://www.taus.net/).

	Tokens			Types			Lemmas		
	SRC	TGT	PE	SRC	TGT	PE	SRC	TGT	PE
EN-DE									
Train (23,000)	384448	403306	411246	18220	27382	31652	10946	21959	25550
Dev (1,000)	17827	19355	19763	2931	3333	3506	1922	2686	2806
Test (2,000)	65120	69812	71483	8061	9765	10502	2626	3976	4282
DE-EN									
Train (25,000)	437833	453096	456163	29745	19866	19172	23532	15422	14131
Dev (1,000)	17578	18130	18313	4426	3583	3642	3589	2828	2836
Test (2,000)	35087	36082	36480	6987	5391	5488	5590	4255	4255

Table 24: Data statistics.

of the WMT APE task. RR measures the repetitiveness inside a text by looking at the rate of non-singleton n-gram types (n=1...4) and combining them using the geometric mean. Larger values indicate a higher text repetitiveness and, as discussed in (Bojar et al., 2016a), suggest a higher chance of learning from the training set correction patterns that are applicable also to the test set. In (Bojar et al., 2016a) we considered the large differences in repetitiveness between APE15 and APE16 data as a possible motivation for the significant baseline improvements achieved by participants in the second round of the task. As we will see in Section 5.4, similar explanations hold for this year's results, in which the higher repetitiveness of English-German data likely contributed to facilitate the task in comparison with the German-English direction.

Table 26 shows, for the same data sets, the Translation Error Rate (TER) (Snover et al., 2006a) and the BLEU score (Papineni et al., 2002) of the original target translations, computed against the human post-edits. In this case, numeric evidence of a higher quality of the original translations can indicate a smaller room for improvement for APE systems (having, at the same time, less to learn during training and less to correct at test stage). On one side, indeed, training on good (or near-perfect) automatic translations can drastically reduce the number of learned correction patterns. On the other side, testing on similarly good translations can drastically reduce the number of corrections required and the applicability of the learned patterns, thus making the task potentially more difficult. Together with the lower repetition rates observed, also the large differences in translation quality between the two APE17 language directions (62.49 BLEU for APE17_EN-DE vs 79.54 for APE17_DE-EN) suggest a higher difficulty for the German-English task. Further indications in this direction are provided by Figures 7 and 8, which plot the TER distribution for the test items in the two data sets. As can be seen, the quality of English-German data is much more balanced compared to German-English, with about 50% of the test items distributed over the first five bins. In particular, what makes a big difference between the two test sets is the proportion of "perfect" test instances having TER=0 (i.e. items that should not be modified by the APE systems). While for English-German they represent 14.0% of the total, for German-English they are about 45.0% of the test data. This means that, for almost half of the German-English test set, any correction made by the APE systems will be unnecessary and penalized by automatic evaluation metrics. This difficult scenario calls for conservative and precise systems able to properly fix errors only in the remaining 50% of the data.

5.1.2 Evaluation metric

System performance was evaluated by computing the distance between *automatic* and *human* post-edits of the machine-translated sentences present in the test set (i.e. for each of the $2,000$ target test sentences). Similar to last year, this distance was measured in terms of TER and BLEU (case-sensitive).[26] Systems were ranked based on the average TER calculated on the test set by using the

[26] In the case of TER, the baseline is computed by averaging the distances between each machine-translated sentence and its human-revised version. The actual evaluation metric is the human-targeted TER (HTER). For the sake of clarity, since TER and HTER compute edit distance in the same way (the only difference is in the origin of the correct sentence used for comparison), henceforth we will use TER to refer to both metrics.

		APE15	APE16	APE17_EN-DE	APE17_DE-EN
RR	SRC	2.905	6.616	7.216	5.225
	TGT	3.312	8.845	9.531	6.841
	PE	3.085	8.245	8.946	6.293

Table 25: Repetition Rate (RR) of the WMT15 (English-Spanish, `news` domain, crowdsourced post-edits), WMT16 (English-German, `IT` domain, professional post-editors), WMT17_EN-DE (English-German, `IT` domain, professional post-editors) and WMT17_DE-EN (German-English, pharmacological domain, professional post-editors) APE task data.

	APE15	APE16	APE17_EN-DE	APE17_DE-EN
TER	23.84	24.76	24.48	15.55
BLEU	n/a	62.11	62.49	79.54

Table 26: Translation quality (TER/BLEU of TGT and proportion of TGTs with TER=0) of the WMT15, WMT16, WMT17_EN-DE and WMT17_DE-EN data.

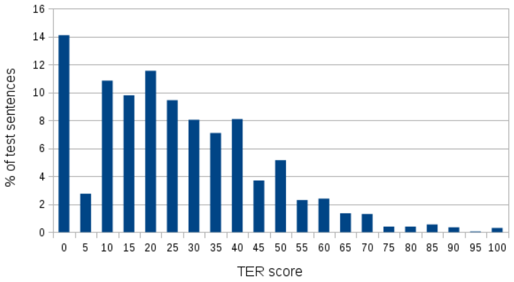


Figure 7: TER distribution over the **EN-DE** test set

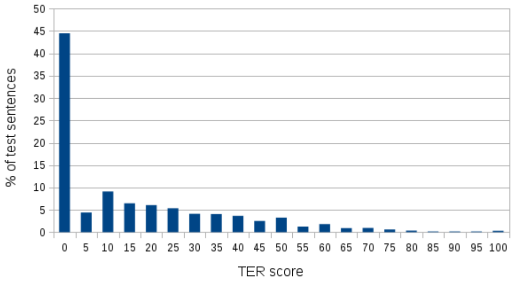


Figure 8: TER distribution over the **DE-EN** test set

TERcom[27] software: lower average TER scores correspond to higher ranks. BLEU was computed using the multi-bleu.perl package[28] available in MOSES.

5.1.3 Baselines

Also this year, the official baseline results are the TER and BLEU scores calculated by comparing the raw MT output with the human post-edits. In practice, the baseline APE system is a *"do-nothing"* system that leaves all the test targets unmodified. Baseline results, the same shown in Table 26, are also reported in Tables 28-29 for comparison with participants' submissions.

In continuity with the previous rounds, we used as additional term of comparison a re-implementation of the method firstly proposed by Simard et al. (2007). It relies on a phrase-based post-editing approach to the task, which represented the common backbone of APE systems before the spread of neural solutions. The system is based on Moses (Koehn et al., 2007); translation and reordering models were estimated following the Moses protocol with default setup using MGIZA++ (Gao and Vogel, 2008) for word alignment. For language modeling we used the KenLM toolkit (Heafield, 2011) for standard n-gram modeling with an n-gram length of 5. Finally, the system was tuned on the development set, optimizing TER/BLEU with Minimum Error Rate Training (Och, 2003). The results of this additional term of comparison are also reported in Tables 28-29.

For each submitted run, the statistical significance of performance differences with respect to the baseline and our re-implementation of Simard et al. (2007) was calculated with the bootstrap test (Koehn, 2004).

5.2 Participants

Seven teams participated in the English-German task by submitting a total of fifteen runs. Two of them also participated in the German-English task with five submitted runs. Participants are listed in Table 27, and a short description of their systems is provided in the following.

Adam Mickiewicz University. AMU's (EN-DE) participation explores and combines multiple neural architectures available in the Marian toolkit.[29] They include single source (either

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

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

[29] `https://github.com/marian-nmt/marian`

ID	Participating team
EN-DE	
AMU	Adam Mickiewicz University, Poland (Junczys-Dowmunt and Grundkiewicz, 2017)
CUNI	Univerzita Karlova v Praze, Czech Republic (Variš and Bojar, 2017)
DCU	Dublin City University, Ireland (Hokamp, 2017)
FBK	Fondazione Bruno Kessler, Italy (Chatterjee et al., 2017)
JXNU	Jiangxi Normal University, Nanchang, China (Tan et al., 2017a)
LIG	University of Lille & University Grenoble, France (Berard et al., 2017)
USAAR	Saarland University, Germany
DE-EN	
FBK	Fondazione Bruno Kessler, Italy (Chatterjee et al., 2017)
LIG	University of Lille & University Grenoble, France (Berard et al., 2017)

Table 27: Participants in the WMT17 Automatic Post-editing task.

$src \rightarrow pe$ or $mt \rightarrow pe$) and multi-source models ($\{src, mt\} \rightarrow pe$), the latter being able to combine information from raw MT output and original source language input. Different attention mechanisms are explored, including soft attention (looking at information anywhere in the source sequence during decoding) and hard monotonic attention (looking at one encoder state at a time from left to right, thus being more conservative and faithful to the original input), which are combined in different ways in the case of multi-source models. The artificial data provided by Junczys-Dowmunt and Grundkiewicz (2016) are used to boost performance by increasing the size of the corpus used for training.

Univerzita Karlova v Praze. CUNI's (EN-DE) system is based on the character-to-character neural network architecture described in (Lee et al., 2016). This architecture was compared with the standard neural network architecture proposed by Bahdanau et al. (2014) which uses byte-pair encoding (Sennrich et al., 2015) for generating translation tokens. During the experiments, two setups have been compared for each architecture: *i)* a single encoder with SRC and MT sentences concatenated, and *ii)* a two-encoder system, where each SRC and MT sentence is fed to a separate encoder. The submitted system uses the two-encoder architecture with a character-level encoder and decoder. The initial state of the decoder is a weighted combination of the final states of the encoders. Attention is computed separately over each encoder. The model was trained using both the WMT17 training data and the artificial data provided by Junczys-Dowmunt and Grundkiewicz (2016). The WMT17 training dataset was sampled to match the size of the artificial data. The submitted primary submission used beam-search for decoding while greedy decoding was used for the contrastive submission.

Dublin City University. DCU's (EN-DE) submission is an ensemble of neural MT systems with different input factors, designed to jointly tackle both the APE task and the Word-Level QE task. Word-Level features which have proven effective for QE, such as word-alignments, part-of-speech tags, and dependency labels, are included as input factors to neural machine translation systems, which are trained to output Post-Edited MT hypotheses. Concatenated *source* + *MT hypothesis* are also used as an input representation for some models. The system makes extensive use of the synthetic training data provided by Junczys-Dowmunt and Grundkiewicz (2016), as well as min-risk training for fine-tuning (Shen et al., 2016). The neural systems, which use different input representations but share the same output vocabulary, are then ensembled together in a log-linear model which is tuned for the TER metric using MERT.

Fondazione Bruno Kessler. FBK's (EN-DE & DE-EN) submission extends the existing NMT implementation in the Nematus toolkit (Sennrich et al., 2016) to train an ensemble of multi-source neural APE systems. Building on previous participations based on the phrase-based paradigm (Chatterjee et al., 2015a, 2016), and similar to (Libovický et al., 2016), such systems jointly learn from source and target information in order to increase robustness and precision of the automatic corrections. The n-best hypotheses produced by this ensemble are further re-ranked using features based on the edit distance between the original MT output and each APE hypothesis, as well as other statistical models (n-gram language model and operation sequence model). For English-German, generic models are trained using the ∼4M synthetic data provided by Junczys-Dowmunt and Grundkiewicz (2016), and then fine-tuned with in-domain data. Similarly, for German-English, synthetic post-editing training data are created by round-trip translation of a sub-set of parallel data released in the medical task at WMT'14 (Bojar et al., 2014).

Jiangxi Normal University. JXNU's (EN-DE) system contains three neural automatic post-editing models: *npe_baseline*, *npe_minor* and *npe_single*. Based on Junczys-Dowmunt and Grundkiewicz (2016), the *npe_baseline* model is created and trained with the training set officially released by the evaluation campaign. The *npe_minor* model is obtained by fine-tuning *npe_baseline* with a triplets corpus including raw machine translation outputs needing four or less edit operations. The *npe_single* model is obtained by fine-tuning *npe_baseline* with a triplets corpus containing machine translations needing at most two edit operations. The output of these three systems is integrated into an n-best list of translations hypotheses, which are scored and ranked by means of a sentence-level QE approach (Specia et al., 2013) and a statistical language model (Stolcke, 2002). Since the raw machine translation outputs can be classified into five grades according to the above sentence-level QE score, the best output can be selected from the n-best list in accordance with the raw MT outputs' grading. The features used by these models can mitigate the over-correction problem emerged in previous rounds of the APE task (Bojar et al., 2016a).

University of Lille & University of Grenoble. LIG's (EN-DE & DE-EN) submission is a neural-based APE system that exploits the approach proposed by Libovický et al. (2016): instead of predicting words, it predicts edit operations (*keep*, *delete*, or *insert* a word). An advantage of this approach, is that it is very easy to learn to replicate the ("*do-nothing*") baseline, by just predicting *keep* operations. By contrast, it can be hard for a classic NMT model to learn the identity function, in particular because of the unknown word problem, and because of the limited amounts of training data. LIG's submission proposes a number of improvements over this method: the simplest model ('*Contrastive-Forced*') uses a task-specific attention mechanism, which forces the decoder to look at the right word in the input (i.e., the word being post-edited). This simple approach gives very good results on the English-German task in limited data conditions. Finally, they also propose a chained architecture ('*Contrastive-Chained*'), which uses two different models (and two different training objectives): a translation model ($src \rightarrow mt$), and a post-editing model ($mt \rightarrow pe$). The attention vectors over src learned by the translation model are used by the post-editing model to give additional contextual information (when predicting a new edit operation, it can look at the mt word to post-edit, and at the src words that are aligned to this word.) This approach is a way to incorporate the source sentence into the proposed framework, and gives promising results on the English-German task, when adding more data ('*primary*' models).

Saarland University. USAAR's (EN-DE) submission combines a neural model and an operation sequence (OSM) phrase-based (Pal et al., 2016c) model. The neural system is trained on a bidirectional (forward-backward) RNN-based encoder-decoder[30] MT model (Bahdanau et al., 2014) trained for $mt \rightarrow pe$ translation. The network has been trained for 5 days using a hyper-parameter setting similar to (Pal et al., 2016b). Training data consists of WMT-2016, 2017 APE data (23K) and 4.5M artificial APE data (Junczys-Dowmunt and Grundkiewicz, 2016). The OSM phrase-based system (Pal et al., 2016c) consists of three basic components: corpus pre-processing, hybrid word alignment (Pal et al., 2016a) and a vanilla setting of a phrase-based MT system integrated with the hybrid word alignment. The model used 23K (*target, human post-edit*) data for training. Experiments on the WMT-2017 test set using both the neural and the OSM-based APE systems revealed that the neural system provides better performance for short sentences (less than 15 words) and the OSM-based APE model performs better for the longer ones. A manual inspection indicates that the neural system suffers from a "lack of coverage" while translating longer sentences. There-

[30] The system used is GroundHog – `https://github.com/lisa-groundhog/GroundHog`.

fore, the final submission was based on a mix of neural translations for short test sentences and OSM translations for the longer ones.

5.3 TER/BLEU results

Participants' TER and BLEU results are shown in Tables 28 (English-German) and 29 (German-English). The submitted runs are ranked based on the average TER (case-sensitive), which is the APE task primary evaluation metric. Overall, similar to last year, TER and BLEU rankings do not show major differences. The main ones can be found in the English-German task where: *i)* two mid-ranked primary submissions (USAAR and JXNU) are inversely ordered by the two metrics, and *ii)* the phrase-based APE (worse in terms of TER) would outperform the "*do-nothing*" strategy by around 0.48 BLEU points. In the German-English task, TER and BLEU rankings differ in the ordering of a primary submission and the "*do-nothing*" baseline, but the negligible score differences are not significant. As we will see in Section5.5, for English-German, the human evaluation based on direct assessment (DA) suggests a third different ranking that is slightly closer to the BLEU-based one (two primary submissions are ranked in the same position, while with TER this happens only in one case). On German-English, a slight preference is confirmed for the BLEU-based ranking as shown by the small difference (0.1) in average DA scores in favour of the "*do-nothing*" baseline over the second-ranked primary submission. However, due to the small differences in systems' architectures and results, it's not surprising that different metrics and evaluation criteria produce slightly different rankings. Also this year, it's hence difficult to draw definite conclusions about which automatic metric is more reliable.

English-German Compared to previous rounds of the APE task, the most noticeable aspect is that this year, for the first time, all participants managed to beat the MT baseline at least with their primary submission.[31] This steady improvement has been mainly driven by the massive migration to the neural approach, which in 2016 allowed the winning system to achieve impressive results (-3.24 TER, +5.54 BLEU with respect to the baseline). This year, the gains on English-German data are even larger, with the winning system scoring -4.88 TER and +7.58 BLEU points better than the MT baseline. The technology advancement is evident if we look at our second term of comparison: the re-implementation of the phrase-based approach by Simard et al. (2007). Last year, on English-German, the results of this method were better than the baseline and in a middle position in the official participants' ranking. This year, on the same language direction, they are almost identical to those achieved in 2016, but also: *i)* worse than the baseline in terms of TER (+0.21), *ii)* slightly better in terms of BLEU (+0.48) and *iii)* competitive only against the contrastive submission of one participant. Considering the distance between the same phrase-based approach and the baseline as an indicator of the task difficulty across different rounds of the task, we hypothesize that the good results achieved by this year's participants are mainly due to improved techniques rather than "easier" test data. Indeed, for English-German where a comparison with last year is possible, the close repetition rate and BLEU scores reported in Tables 25 and 26 reveal a similar level of difficulty for the APE16 and APE_17 test data.

ID	Avg. TER	BLEU
FBK Primary	19.6	70.07
AMU Primary	19.77	69.5
AMU Contrastive	19.83	69.38
DCU Primary	20.11	69.19
DCU Contrastive	20.25	69.33
FBK Contrastive	20.3	69.11
FBK_USAAR Contr.	21.55	67.28
USAAR Primary	23.05	65.01
LIG Primary	23.22	65.12
JXNU Primary	23.31	65.66
LIG Contrastive-Forced	23.51	64.52
LIG Contrastive-Chained	23.66	64.46
CUNI Primary	24.03†	64.28
USAAR Contrastive	24.17	63.55
Baseline	24.48	62.49
(Simard et al., 2007)	24.69	62.97
CUNI Contrastive	25.94	61.65

Table 28: Results for the WMT17 APE **EN-DE** task – average TER (↓), BLEU score (↑). The † indicates a difference from the MT baseline that is not statistically significant.

German-English On German-English, the improvements of the top submission over the baseline are smaller (-0.26 TER, +0.28 BLEU) but still statistically significant. Such smaller gains, ob-

[31] In 2015, none of the submitted runs were able to consistently improve over the raw MT output. Last year, only half of the runs outperformed this baseline.

ID	Avg. TER	BLEU
FBK Primary	15.29	79.82
FBK Contrastive	15.31	79.64†
LIG Primary	15.53†	79.49†
Baseline	15.55	79.54
LIG Contrastive-Forced	15.62†	79.48†
LIG Contrastive-Chained	15.68	79.35
(Simard et al., 2007)	15.74	79.28†

Table 29: Results for the WMT17 APE **DE-EN** task – average TER (↓), BLEU score (↑). The † indicates differences from the MT baseline that are not statistically significant.

tained by systems based on the same approaches adopted for the English-German task, confirm our initial expectations about the different level of difficulty of the two language directions. The interaction between low repetition rates and high translation quality, which certainly played a role in reducing the gap between the primary submissions and the "*do-nothing*" MT baseline, is hence an interesting aspect for more thorough explorations in future rounds of the APE task. Also in this case, however, the lowest results achieved by the phrase-based APE baseline (with both metrics) confirm that the switch to neural methods represents a technology advancement in the right direction.

5.4 System/performance analysis

Although all participants built their systems under the same general neural paradigm, results' distribution in a 4.5 TER (and 6.5 BLEU) points interval suggests differences in systems' behaviour that it is worth to explore further. To this aim, and as a complement to global TER/BLEU scores, also this year we performed a more fine-grained analysis of the changes made by each system to the test instances.

5.4.1 Macro indicators: modified, improved and deteriorated sentences

Tables 30 and 31 show the number of modified, improved and deteriorated sentences, respectively for the English-German and the German-English tasks. It's worth noting that, as in the previous rounds and for both language directions, the number of sentences modified by each system is higher than the sum of the improved and the deteriorated ones. This difference is represented by modified sentences for which the corrections do not yield TER variations. This grey area, for which quality improvement/degradation can not be automatically assessed, contributes to motivate the human evaluation discussed in Section 5.5

English-German. As expected, differently from last year where the amount of test sentences modified by the participants had a much larger variance due to the different approaches applied, this year the top English-German systems show a quite homogeneous behaviour. In 2016, out of 11 submitted runs, the number of sentences modified by the top 3 primary submissions (the best one being neural and the others being phrase-based) ranged between 421 and 1,613 (respectively 21.0% and 80.6% of the total). This year, out of 15 submitted runs (all neural-based), the top 3 primary submissions have a number of modified sentences that falls in a much smaller range between 1,583 and 1,607 (between 79.1% and 80.0% of the total). The same holds for systems' precision (i.e. the proportion of improved sentences out of the total amount of modified test items). The top 3 primary submissions, indeed, have a precision ranging in a two points interval from 63.6% to 65.6%, while last year the proportion for the top 3 primary runs was more spread in a 11 points interval from 57.9% to 68.8%. Overall, lower ranked systems show a tendency to either modify less sentences (all submissions with less than 1,000 modified sentences are in the bottom half of the ranking), or to do it with lower precision (all submissions with less than 60.0% precision are in the bottom half of the ranking), or a combination of the two, as in the case of the phrase-based approach (Simard et al., 2007), which is the second less aggressive method and by far the less precise one. In general, looking at system precision numbers, it's worth noting that the close results between the top submissions still leave large room for improvement. Indeed, in the case of the best systems, more than 30 points in precision represent a huge gap to be filled before considering APE a solved problem.

German-English. In this case, the higher difficulty of the task (due to lower repetition rate and higher translation quality, as discussed in Section 5.1.1) changes the global picture provided by our macro indicators. Although the two participating systems were developed under the neural paradigm, their different behaviour is evident from the amount of modified sentences: the two primary submissions respectively modified 270

Systems	Modified	Improved	Deteriorated
FBK Primary	1,607 (80.3%)	1,035 (64.4%)	334 (20.7%)
AMU Primary	1,583 (79.1%)	1,040 (65.6%)	322 (20.3%)
AMU Contrastive	1,583 (79.1%)	1,044 (65.9%)	326 (20.5%)
DCU Primary	1,592 (79.6%)	1,014 (63.6%)	361 (22.6%)
DCU Contrastive	1,558 (77.9%)	1,012 (64.9%)	329 (21.1%)
FBK Contrastive	1,597 (79.8%)	996 (62.3%)	344 (21.5%)
FBK_USAAR Contrastive	1,675 (83.7%)	920 (55.0%)	482 (28.7%)
USAAR Primary	744 (37.2%)	461 (61.9%)	160 (21.5%)
LIG Primary	1,168 (58.4%)	629 (53.8%)	306 (26.1%)
JXNU Primary	1,385 (69.2%)	678 (48.9%)	404 (29.1%)
LIG Contrastive-Forced	719 (35.9%)	412 (57.3%)	166 (23.1%)
LIG Contrastive-Chained	814 (40.7%)	422 (51.8%)	217 (26.6%)
CUNI Primary	1,513 (75.6%)	713 (47.1%)	515 (34.0%)
USAAR Contrastive	306 (15.3%)	179 (58.4%)	76 (24.8%)
(Simard et al., 2007)	571 (28.5%)	211 (36.9%))	244 (42.7%)
CUNI Contrastive	1577 (78.8%)	644 (40.8%)	663 (42.0%)

Table 30: Number of test sentences modified, improved and deteriorated by each run submitted to the **EN-DE** task.

Systems	Modified	Improved	Deteriorated
FBK Primary	270 (13.5%)	108 (40.0%)	78 (28.9%)
FBK Contrastive	364 (18.2)	135 (37.0%	118 (32.4%)
LIG Primary	64 (3.2%)	27 (42.1%)	24 (37.5%)
LIG Contrastive-Forced	47 (2.3%)	13 (27.6%)	21 (44.7%)
LIG Contrastive-Chained	64 (3.2%)	27 (42.1%)	46 (71.9%)
(Simard et al., 2007)	139 (6.9%)	30 (21.6%)	69 (49.6%)

Table 31: Number of test sentences modified, improved and deteriorated by each run submitted to the **DE-EN** task.

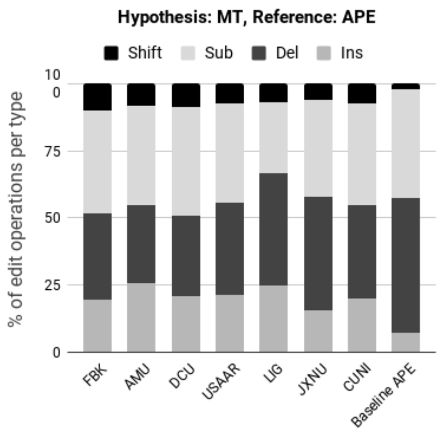


Figure 9: System behaviour (primary submissions) for **EN-DE** – TER(MT, APE)

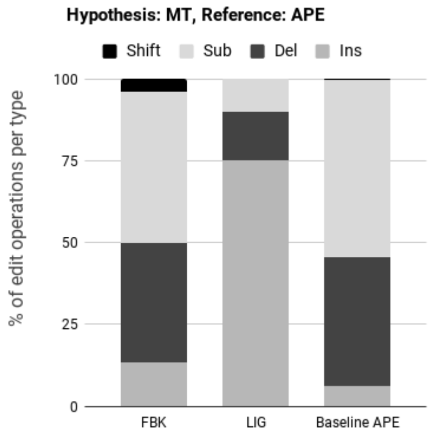


Figure 10: System behaviour (primary submissions) for **DE-EN** – TER(MT, APE)

(13.5%) and 64 (3.2%) test items. On one side, the small number of modified sentences compared to English-German indicates systems' ability to keep under control the number of unnecessary corrections. If we consider that almost half of the test items are "perfect" translations that should be kept unchanged (see Table26), a rather conservative approach is indeed a desired behaviour. On the other side, however, precision scores are much lower compared to those observed in the English-German task. Even for an "easy" target language like English, coping with data featuring low repetition rates and high translation quality is hence a still open challenge.

5.4.2 Micro indicators: edit operations

Also this year we performed a more fine-grained analysis of systems' behaviour in order to discover possible differences in the way they correct the test set instances. To this aim, we looked at the distribution of the edit operations done by each system (insertions, deletions, substitutions and shifts) by computing the TER between the original MT output and the output of each system taken as reference (only for the primary submissions). The outcomes of this analysis are shown in Figures 9 (English-German) and 10 (German-English).

English-German. As expected, compared to last year, the plot in Figure 9 does not show large differences between similar neural-based submissions. All of them are characterized by a rather homogeneous distribution of the types of correction patterns applied, with a slight dominance of substitutions for the top submissions (between 37.0% and 40.0%) and a slight dominance of deletions for the others (between 34.5% and 42.1%). Another quite visible correlation is the one between shift operations and performance results, which tend to decrease for systems that perform less reordering (also last year, the winning neural system had a significantly larger amount of shifts compared to the others). Interestingly, also in this case the phrase-based baseline (the weakest APE system in terms of results) is a clear outlier. It performs the lowest number of shifts (2.2% vs 9.7% of the top submission), the lowest number of insertions (7.1% vs 19.5%) and the largest number of deletions (50.2% vs 32.1%). This indicates a scarce capability of the phrase-based approach to learn reordering rules and its tendency to replace them with more radical deletion operations.

German-English. As shown by Figure 10, the two primary submissions for this task have a quite different behaviour. In addition to the large differences in the number of modified, improved and deteriorated sentences (see Table 31), the distribution of the edit operations performed on test data indicates opposite strategies. Also in this case, the distribution is more homogeneous for the best performing system, with a dominance of substitutions and around 4.0% of shifts (though less than in the English-German task, where they were around 10.0%). The second system has a much more unbalanced distribution, with lots of insertions and no shifts in the few sentence corrections it returned. The distribution for the phrase-based APE baseline is more similar to the best system but, as shown in Table 31, its corrections are by far the less reliable ones. Apart from these considerations, it is hard to draw clear conclusions since the different correction strategies of the three methods result in close final scores. Indeed, as shown in Table 29, only 0.24 TER and 0.33 BLEU points separate the two primary systems, while 0.45 TER and 0.54 BLEU points separate the best system from the phrase-based baseline. The small improvements of the primary submissions over the "*do-nothing*" MT baseline suggest that, independently from the different correction strategies applied, both primary submissions definitely suffered from the large amount of "perfect" translation in the test set (around 45.0%). However, while automatic evaluation metrics like TER and BLEU always penalize unnecessary corrections of good translations, there is a chance that some of these corrections are acceptable paraphrases rather than sentence deteriorations. One of the objectives of the human evaluation discussed in the next section is to check if this phenomenon has a visible impact on performance.

5.5 Human evaluation

To assess the quality of the output of the APE systems and produce a ranking based on human judgment, as well as analyze how humans perceive TER/BLEU performance differences between the submitted systems, a human evaluation of the quality of automatic post-edits was carried out using Direct Assessment (DA) (Graham et al., 2013, 2016). Since sufficient crowd-sourced workers are available for assessing English on Mechanical Turk, the DA evaluation for German to English was completed via quality-controlled crowdsourcing. For English to German, DA judgments were provided by 10 native German speakers from Saarland University, studying language technologies and translation. This subsection describes the human evaluation procedure and presents the results of the evaluation of participants' primary submissions.

5.5.1 Evaluation procedure

Direct Assessment, which is described in more detail in Section 3, elicits human assessments of translation adequacy on an analogue rating scale

Language Pair	EN-DE	DE-EN
# Systems	9	4
# Segs	2,000	2,000
# Total Segs	18,000	8,000
# Unique Segs	9,767	3,415
Overall Saving	46%	57%

Table 32: Total segments prior to sampling for manual evaluation and savings made by combining identical segments (Segs) produced by multiple APE systems.

(0–100), where human assessors are asked to rate how adequately the APE system output expresses the meaning of the human reference translation. DA scores for systems and segments have been shown to be highly repeatable in self-replication experiments (Graham et al., 2015). Thus, DA overcomes the previous challenges associated with lack of reliability of human assessment of MT.

Since we also have a human post-edit available for each MT output in the test set, to make DA outcomes more informative we also included the human post-edits as a hidden system in the evaluation, which will provide some insight into an achievable DA score for a potential system that achieved human-quality post-editing. Additionally, we included the original MT output without any post-editing as a hidden system to discover the baseline DA score for each language pair.

When running the APE manual evaluation, it was possible in many cases to take advantage of the fact that multiple systems can produce identical outputs, as was begun in evaluation of the News task in WMT15 (Bojar et al., 2015). Table 32 shows numbers of translations in total for all APE systems, as well as savings in terms of annotation effort that was gained by combining identical system outputs prior to running the evaluation, where, as expected, a substantial saving was made due to the fact that the systems quite often produced the same output. In terms of human effort involved in carrying out the manual evaluation, Table 33 shows numbers of judgments collected in total for each language pair and number of assessments contributing to the final DA score for APE systems on average.

When carrying out a manual evaluation of any kind, it is important to consider the consistency of annotators with the aim of estimating, where the evaluation to be repeated, how likely it would be that the same conclusions would be drawn.

	Systems	**Assess**	**Assess/Sys**
EN-DE	9	11,492	1,277
DE-EN	4	7,193	1,798

Table 33: Amount of data (assessments after “de-collapsing” *multi-system outputs*) collected in the WMT17 APE manual evaluation campaign and numbers of assessments per system.

When an analogue scale is employed for human assessment, consistency of human assessors cannot be evaluated in the usual way, such as the Kappa coefficient, commonly employed for evaluating the consistency of human assessors when discrete quality judgments or relative preference judgments are collected. Instead, for analogue scale data, we examine the consistency of individual human assessors according to their ability to discriminate between the quality of pairs of known worse quality translations, known as bad reference pairs, where original translations produced by the APE systems are degraded automatically. In addition, repeat assessments of the same translation are given to human assessors to see how reliably they assign similar scores to similar quality translations. Hiding bad reference and repeat translation pairs within hits allows a significance test to be carried out for each human judge investigating if their score distributions show a significant difference where there should be one, and another test to check that no significant difference shows up for repeated assessment of the same translation.

As such, proportions of human assessors and whether they discriminate between the quality of bad reference pairs and repeat translations are shown in Table 34. Notably, all of the student translators (for EN-DE) passed the DA's quality control mechanism by assigning significantly lower scores to degraded translations, while 54%, a usual number of crowd-sourced workers (for DE-EN), passed quality control.

Proportions of workers showing a non significant difference in repeat items at first appears lower than usual for DA, at 91% for EN-DE and 93% for DE-EN, as this proportion has been between 97 and 100% for DA in past evaluations. However, on closer inspection, the total number of assessors showing a significant difference for repeat items is as low as three assessors and proportions are therefore exaggerated due to the low number of workers involved in the evaluation overall.

	All	(A) Sig. Diff. Bad Ref.	(A) & No Sig. Diff. Exact Rep.
EN-DE	11	11 (100%)	10 (91%)
DE-EN	54	29 (54%)	27 (93%)

Table 34: Number of unique Mechanical Turk workers, (A) those whose scores for bad reference pairs were significantly different and numbers of unique human assessors in (A) whose scores for exact repeat assessments also showed no significant difference.

Prior to computing final DA scores for systems, in order to iron out differences in scoring strategies of distinct human assessors, human assessment scores for translations were first standardized according to each individual human assessor's overall mean and standard deviation score. Average standardized scores for individual segments belonging to a given system are then computed, before the final overall DA score for that system is computed as the average of its segment scores.

5.5.2 Human evaluation results

Table 35 includes DA results for English-German and Table 36 shows results for German-English APE systems. 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.

#	Ave %	Ave z	System
–	84.8	0.520	HUMAN POST_EDIT
1	78.2	0.261	AMU
	77.9	0.261	FBK
	76.8	0.221	DCU
4	73.8	0.115	JXNU
5	71.9	0.038	USAAR
	71.1	0.014	CUNI
	70.2	−0.020	LIG
–	68.6	−0.083	NO POST_EDIT

Table 35: EN-DE DA Human evaluation results showing average raw DA scores (Ave %) and average standardized scores (Ave z), lines between systems indicate clusters according to Wilcoxon rank-sum test at p-level $p \leq 0.05$.

#	Ave %	Ave z	System
–	81.9	0.199	HUMAN POST_EDIT
1	76.8	0.040	FBK
	75.3	−0.007	LIG
	75.4	−0.008	NO POST_EDIT

Table 36: DE-EN DA Human evaluation results showing average raw DA scores (Ave %) and average standardized scores (Ave z), lines between systems indicate clusters according to Wilcoxon rank-sum test at p-level $p \leq 0.05$.

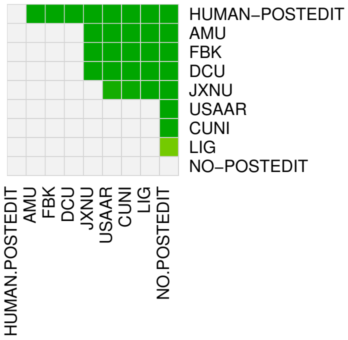


Figure 11: EN-DE Wilcoxon rank-sum significance test results for pairs of systems competing in the APE task, where a green cell denotes a significant win for the system in a given row over the system in a given column, at $p \leq 0.05$.

Figures 11 and 12 show head to head significance test results for English-German and German-English systems participating in the APE task, as well as the two additional "systems" where either no post-editing or human post-editing was carried out, where a darker shade of green signifies a lower p-value and a conclusion made with more certainty.

English-German. For this language direction, the ranking produced by DA is slightly different from those based on TER/BLEU. This is not surprising if we consider the close performance results measured with automatic metrics. With primary submissions compressed in a relatively small TER/BLEU interval, different system orders are in fact likely to emerge also from manual evalua-

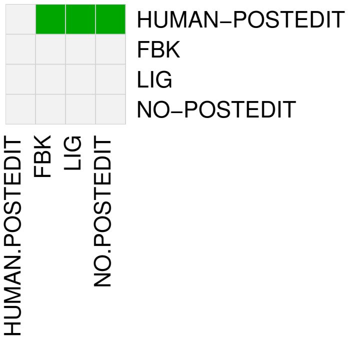


Figure 12: DE-EN Wilcoxon rank-sum significance test results for pairs of systems competing in the APE task, where a green cell denotes a significant win for the system in a given row over the system in a given column, at $p \leq 0.05$.

tion. Overall, as shown in Table 35, three systems emerge as significantly better than the others. This ranking is comparable to the one obtained with automatic metrics, although the top two systems (FBK and AMU) are switched, but this is in-line with the human evaluation that showed no significant difference between the two. This is also in-line with TER/BLEU rankings, for which the three systems are the only primary systems with TER<20.00 and BLEU>69.00. In agreement with the BLEU-based ranking, the JXNU submission ranks in fourth position in its own cluster. This represents the main difference with the TER-based ranking (in which it occupies the 6th place), which suggests a higher agreement between DA and BLEU. The remaining three systems, which feature rather close TER/BLEU scores, are positioned in the same lower cluster, though in a different order, again with small raw DA score differences.

Apart from these general considerations, which are difficult to project into conclusive indications about the reliability of our two automatic metrics, two major outcomes are evident. First, the technology advancement with respect to the 2016 round is also confirmed by DA scores, which indicate that all the systems are significantly better than the "*do-nothing*" baseline (NO POST_EDIT). Last year, in contrast, all participants but one were in the same cluster of the baseline. The downside is that, despite the significant progress made, APE systems are still far from human quality. Average DA scores indicate that the distance between the top primary submissions and human post-edits is in fact similar to the distance that separates them from the primary submissions in the bottom cluster.

German-English. Also DA scores confirm the higher difficulty of the German-English task. As expected, also in this case human quality is much higher, with a gap that is even larger compared to the distance observed in Figure 35. Moreover, while in terms of automatic metrics the improvement over the baseline for the top ranked system was statistically significant, the DA-based ranking places the two primary systems in the same cluster of the baseline.

5.6 Lessons learned and outlook

The third round of the APE task has marked a further step forward from the previous ones both in terms of participants (one more than in 2016) and, most importantly, in terms of the deployed technology. Concerning the latter aspect, the wide adoption of neural approaches has led, for the first time, to significant improvements over the baselines for all participants. On English-German data we observed the largest gains, which are up to -4.9 TER and +7.6 BLEU points for the top submission. On German-English, a more difficult task due to lower repetition rate and higher translation quality of the test data, the improvements of the top submission over the baseline are smaller (-0.26 TER, +0.28 BLEU) but still statistically significant. With respect to previous years, similar design and training choices (e.g. the use of multi-source solutions and additional synthetic training data), produced a more compact ranking of the participating systems but, at the same time, resulted in submissions that still feature different behaviour that deserve closer inspection in future.

Despite the technology improvement, some major challenges are still open. The main one is how to better handle the difficult case in which an automatic translation is already (or near-) perfect and APE systems should abstain from performing useless (or risky) corrections. Another limitation of current solutions is their inefficacy in generalizing the learned correction patterns, so that training data featuring low repetitiveness can be better exploited to learn useful correction patterns.

From the performance evaluation standpoint, the selection of the best metric is still debatable. TER (the official one in all the APE rounds so far) and BLEU produce slightly different rankings, which both differ from those produced by human evaluation with direct assessment. The comparison with DA indicates a small preference for the BLEU-based ranking, but drawing definite conclusions about the suitability of the two metrics is difficult due to the small performance differences observed. Most likely, future rounds of the task will hence keep the the evaluation setting unaltered, possibly focusing on the aforementioned challenges to increase the level of difficulty and further raise the interest on the APE problem.

Acknowledgments

This work was supported in parts by the QT21, QTLeap, EXPERT and CRACKER projects funded by the European Commission (7th Framework Programme and H2020). Further datasets were donated by University of Helsinki and Yandex. The APE task organizers would also like to thank Text&Form for producing the manual post-edits and the annotators involved in the manual evaluation.

References

Hervé Abdi. 2007. The bonferroni and šidák corrections for multiple comparisons. *Encyclopedia of measurement and statistics*, 3:103–107.

Eleftherios Avramidis. 2017a. Comparative quality estimation for machine translation: Observations on machine learning and features. *The Prague Bulletin of Mathematical Linguistics (PBML)*, 108:307–318.

Eleftherios Avramidis. 2017b. Sentence-level quality estimation by predicting HTER as a multi-component metric. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Alexandre Berard, Laurent Besacier, and Olivier Pietquin. 2017. LIG-CRIStAL Submission for the WMT 2017 Automatic Post-Editing Task. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Ergun Biçici. 2017. Predicting Translation Performance with Referential Translation Machines. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Frédéric Blain, Varvara Logacheva, and Lucia Specia. 2016. Phrase level segmentation and labelling of machine translation errors. *LREC16*.

Frédéric Blain, Carolina Scarton, and Lucia Specia. 2017. Bilexical Embeddings for Quality Estimation. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Ondrej 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, Yvette Graham, and Amir Kamran. 2017a. Results of the WMT17 Metrics Shared Task. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Jindřich Helcl, Tom Kocmi, Jindřich Libovický, and Tomáš Musil. 2017b. Results of the WMT17 Neural MT Training Task. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. 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. In press.

Franck Burlot, Pooyan Safari, Matthieu Labeau, Alexandre Allauzen, and François Yvon. 2017. LIMSI@WMT'17. 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.

Rajen Chatterjee, M. Amin Farajian, Matteo Negri, Marco Turchi, Ankit Srivastava, and Santanu Pal. 2017. Multi-source Neural Automatic Post-Editing: FBK's participation in the WMT 2017 APE shared task. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Rajen Chatterjee, José G. C. de Souza, Matteo Negri, and Marco Turchi. 2016. The FBK Participation in the WMT 2016 Automatic Post-editing Shared Task. In *Proceedings of the 11th Workshop on Statistical Machine Translation (WMT)*.

Rajen Chatterjee, Turchi Turchi, and Matteo Negri. 2015a. The FBK Participation in the WMT15 Automatic Post-editing Shared Task. In *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT)*.

Rajen Chatterjee, Marion Weller, Matteo Negri, and Marco Turchi. 2015b. Exploring the Planet of the APEs: a Comparative Study of State-of-the-art Methods for MT Automatic Post-Editing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics)*, Beijing, China.

Zhiming Chen, Yiming Tan, Chenlin Zhang, Qingyu Xiang, Lilin Zhang, Maoxi Li, and Mingwen WANG. 2017. Improving Machine Translation Quality Estimation with Neural Network Features. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Yongchao Deng, Jungi Kim, Guillaume Klein, Catherine KOBUS, Natalia Segal, Christophe Servan, Bo Wang, Dakun Zhang, Josep Crego, and Jean Senellart. 2017. SYSTRAN Purely Neural MT Engines for WMT2017. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Mattia Antonino Di Gangi, Nicola Bertoldi, and Marcello Federico. 2017. FBK's Participation to the English-to-German News Translation Task of WMT 2017. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Shuoyang Ding, Huda Khayrallah, Philipp Koehn, Matt Post, Gaurav Kumar, and Kevin Duh. 2017. The JHU Machine Translation Systems for WMT 2017. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

John Duchi and Yoram Singer. 2009. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10:2899–2934.

Melania Duma and Wolfgang Menzel. 2017. UHH Submission to the WMT17 Quality Estimation Shared Task. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Carlos Escolano, Marta R. Costa-jussà, and José A. R. Fonollosa. 2017. The TALP-UPC Neural Machine Translation System for German/Finnish-English Using the Inverse Direction Model in Rescoring. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. 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.

Qin Gao and Stephan Vogel. 2008. Parallel Implementations of Word Alignment Tool. In *Proceedings of the ACL 2008 Software Engineering, Testing, and Quality Assurance Workshop*, pages 49–57, Columbus, Ohio.

Mercedes García-Martínez, Ozan Caglayan, Walid Aransa, Adrien Bardet, Fethi Bougares, and Loïc Barrault. 2017. LIUM Machine Translation Systems for WMT17 News Translation Task. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2014. Is machine translation getting better over time? In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 443–451, Gothenburg, Sweden. Association for Computational Linguistics.

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

Yvette Graham, Nitika Mathur, and Timothy Baldwin. 2015. Accurate evaluation of segment-level machine translation metrics. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies*, Denver, Colorado.

Stig-Arne Grönroos, Sami Virpioja, and Mikko Kurimo. 2017. Extending hybrid word-character neural machine translation with multi-task learning of morphological analysis. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. 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.

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

Ralf Herbrich, Tom Minka, and Thore Graepel. 2007. Trueskill$^{\text{TM}}$: a bayesian skill rating system. In *Advances in neural information processing systems*, pages 569–576.

Teresa Herrmann, Jan Niehues, and Alex Waibel. 2013. Combining Word Reordering Methods on different Linguistic Abstraction Levels for Statistical Machine Translation. In *Proceedings of the Seventh Workshop on Syntax, Semantics and Structure in Statistical Translation*, Altanta, Georgia, USA.

Chris Hokamp. 2017. Ensembling Factored Neural Machine Translation Models for Automatic Post-Editing and Quality Estimation. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Chester Holtz, Chuyang Ke, and Daniel Gildea. 2017. University of Rochester WMT 2017 NMT System Submission. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Matthias Huck, Fabienne Braune, and Alexander Fraser. 2017. 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.

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.

Antonio Jimeno Yepes, Aurelie Neveol, Mariana Neves, Karin Verspoor, Ondrej Bojar, Arthur Boyer, Cristian Grozea, Barry Haddow, Madeleine Kittner, Yvonne Lichtblau, Pavel Pecina, Roland Roller, Rudolf Rosa, Amy Siu, Philippe Thomas, and Saskia Trescher. 2017. Findings of the WMT 2017 Biomedical Translation Shared Task. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Log-linear combinations of monolingual and bilingual neural machine translation models for automatic post-editing. In *Proceedings of the First Conference on Machine Translation*, pages 751–758, Berlin, Germany. Association for Computational Linguistics.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2017. An Exploration of Neural Sequence-to-Sequence Architecturesfor Automatic Post-Editing. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark.

Hyun Kim and Jong-Hyeok Lee. 2016. Recurrent neural network based translation quality estimation. In *Proceedings of the First Conference on Machine Translation*, pages 787–792, Berlin, Germany. Association for Computational Linguistics.

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-Estimator using Multilevel Task Learning with Stack Propagation for Neural Quality Estimation. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

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

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej 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.

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.

Julia Kreutzer, Shigehiko Schamoni, and Stefan Riezler. 2015. Quality estimation from scratch (quetch): Deep learning for word-level translation quality estimation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 316–322, Lisbon, Portugal. Association for Computational Linguistics.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2016. Fully character-level neural machine translation without explicit segmentation. *arXiv preprint arXiv:1610.03017*.

Jindřich Libovický, Jindřich Helcl, Marek Tlustý, Ondřej Bojar, and Pavel Pecina. 2016. CUNI at Post-editing and Multimodal Translation Tasks. In *Proceedings of the 11th Workshop on Statistical Machine Translation (WMT)*, Berlin, Germany.

Varvara Logacheva, Chris Hokamp, and Lucia Specia. 2016. MARMOT: A Toolkit for Translation Quality Estimation at the Word Level. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference*, Portorož, Slovenia.

Ngoc Quang Luong, Laurent Besacier, and Benjamin Lecouteux. 2014. Lig system for word level qe task at wmt14. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 335–341, Baltimore, Maryland, USA. Association for Computational Linguistics.

Pranava Swaroop Madhyastha, Xavier Carreras, and Ariadna Quattoni. 2014. Learning task-specific bilexical embeddings. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 161–171, Dublin, Ireland.

André Martins, Marcin Junczys-Dowmunt, Fabio Kepler, Ramn Astudillo, Chris Hokamp, and Roman Grundkiewicz. 2017a. Pushing the limits of translation quality estimation. *Transactions of the Association for Computational Linguistics*, 5:205–218.

André F. T. Martins, Ramón Astudillo, Chris Hokamp, and Fabio Kepler. 2016. Unbabel's participation in the wmt16 word-level translation quality estimation shared task. In *Proceedings of the First Conference on Machine Translation*, pages 806–811, Berlin, Germany. Association for Computational Linguistics.

André F. T. Martins, Fabio Kepler, and Jose Monteiro. 2017b. Unbabel's Participation in the WMT17 Translation Quality Estimation Shared Task. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, ACL '03, pages 160–167, Sapporo, Japan.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of Conditional Random Fields.

Robert Östling, Yves Scherrer, Jörg Tiedemann, Gongbo Tang, and Tommi Nieminen. 2017. The Helsinki Neural Machine Translation System. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Gustavo Paetzold and Lucia Specia. 2017. Feature-Enriched Character-Level Convolutions for Text Regression. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Santanu Pal, Sudip Kumar Naskar, and Josef van Genabith. 2016a. Multi-engine and multi-alignment based automatic post-editing and its impact on translation productivity. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2559–2570, Osaka, Japan. The COLING 2016 Organizing Committee.

Santanu Pal, Sudip Kumar Naskar, Mihaela Vela, and Josef van Genabith. 2016b. A Neural Network based Approach to Automatic Post-Editing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 281–286, Berlin, Germany.

Santanu Pal, Marcos Zampieri, and Josef van Genabith. 2016c. Usaar: An operation sequential model for automatic statistical post-editing. In *Proceedings of the First Conference on Machine Translation*, pages 759–763, Berlin, Germany.

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

Raj Nath Patel and Sasikumar M. 2016. Translation quality estimation using recurrent neural network. In *Proceedings of the First Conference on Machine Translation*, pages 819–824, Berlin, Germany. Association for Computational Linguistics.

Jan-Thorsten Peter, Andreas Guta, Tamer Alkhouli, Parnia Bahar, Jan Rosendahl, Nick Rossenbach, Miguel Graça, and Hermann Ney. 2017a. The RWTH Aachen University English-German and German-English Machine Translation System for WMT 2017. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Jan-Thorsten Peter, Hermann Ney, Ondřej Bojar, Ngoc-Quan Pham, Jan Niehues, Alex Waibel, Franck Burlot, François Yvon, Mārcis Pinnis, Valters Sics, Joost Bastings, Miguel Rios, Wilker Aziz, Philip Williams, Frédéric Blain, and Lucia Specia. 2017b. The QT21 Combined Machine Translation System for English to Latvian. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Ngoc-Quan Pham, Jan Niehues, Thanh-Le Ha, Eunah Cho, Matthias Sperber, and Alexander Waibel. 2017. The Karlsruhe Institute of Technology Systems for the News Translation Task in WMT 2017. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Mārcis Pinnis, Rihards Krišlauks, Toms Miks, Daiga Deksne, and Valters Šics. 2017. Tilde's Machine Translation Systems for WMT 2017. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Sylvain Raybaud, David Langlois, and Kamel Smali. 2011. this sentence is wrong. detecting errors in machine-translated sentences. *Machine Translation*, 25(1):1–34.

Matīss Rikters, Chantal Amrhein, Maksym Del, and Mark Fishel. 2017. C-3MA: Tartu-Riga-Zurich Translation Systems for WMT17. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2014. Efficient elicitation of annotations for human evaluation of machine translation. In *Proceedings of the Workshop on Statistical Machine Translation*, Baltimore, Maryland. Association for Computational Linguistics.

Carolina Scarton, Daniel Beck, Kashif Shah, Karin Sim Smith, and Lucia Specia. 2016. Word embeddings and discourse information for quality estimation. In *Proceedings of the First Conference on Machine Translation*, pages 831–837, Berlin, Germany. Association for Computational Linguistics.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for wmt 16. In *Proceedings of the First Conference on Machine Translation*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1692, Berlin, Germany.

Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007. Statistical phrase-based post-editing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 508–515, Rochester, New York. Association for Computational Linguistics.

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

Matthew Snover, Bonnie J. Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006b. A study of translation edit rate with targeted human annotation. In *5th Conference of the Association for Machine Translation in the Americas (AMTA)*, Boston, Massachusetts.

Artem Sokolov, Julia Kreutzer, Kellen Sunderland, Pavel Danchenko, Witold Szymaniak, Hagen Fürstenau, and Stefan Riezler. 2017. A Shared Task on Bandit Learning for Machine Translation. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Kim Harris, Aljoscha Burchardt, Viviven Macketanz, Inguna Skadiņa, Marco Turchi, and Matteo Negri. 2017a. Translation quality and productivity: A study on rich morphology languages. In *Proceedings of the 16th Machine Translation Summit*, Nagoya, Japan.

Lucia Specia, Kim Harris, Aljoscha Burchardt, Marco Turchi, Matteo Negri, and Inguna Skadina. 2017b. Translation Quality and Productivity: A Study on Rich Morphology Languages. In *Proceedings of the 16th Machine Translation Summit*, Nagoya, Japan.

Lucia Specia, Gustavo Paetzold, and Carolina Scarton. 2015. Multi-level Translation Quality Prediction with QuEst++. In *53rd Annual Meeting of the Association for Computational Linguistics and Seventh International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing: System Demonstrations*, pages 115–120, Beijing, China.

Lucia Specia, Kashif Shah, Jose G.C. de Souza, and Trevor Cohn. 2013. Quest - a translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 79–84, Sofia, Bulgaria.

A. Stolcke. 2002. Srilm - an extensible language modeling toolkit. In *ICSLP*, pages 901–904, Denver, CO.

Roman Sudarikov, David Mareček, Tom Kocmi, Dušan Variš, and Ondřej Bojar. 2017. CUNI Submission in WMT17: Chimera Goes Neural. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. 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.

Yiming Tan, Zhiming Chen, Liu Huang, Lilin Zhang, Maoxi Li, and Mingwen Wang. 2017a. Neural Post-Editing Based on Quality Estimation. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Zhixing Tan, Boli Wang, Jinming Hu, Yidong Chen, and xiaodong shi. 2017b. XMU Neural Machine Translation Systems for WMT 17. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Long Trieu, Trung-Tin Pham, and Le-Minh Nguyen. 2017. The JAIST Machine Translation Systems for WMT 17. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Dušan Variš and Ondřej Bojar. 2017. CUNI System for WMT17 Automatic Post-Editing Task. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Yuguang Wang, Shanbo Cheng, Liyang Jiang, Jiajun Yang, Wei Chen, Muze Li, Lin Shi, Yanfeng Wang, and Hongtao Yang. 2017. Sogou Neural Machine Translation Systems for WMT17. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Krzysztof Wolk and Krzysztof Marasek. 2017. PJIIT's systems for WMT 2017 Conference. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Jia Xu, Yi Zong Kuang, Shondell Baijoo, Jacob Hyun Lee, Uman Shahzad, Mir Ahmed, Meredith Lancaster, and Chris Carlan. 2017. Hunter MT: A Course for Young Researchers in WMT17. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Alexander Yeh. 2000. More Accurate Tests for the Statistical Significance of Result Differences. In *Coling-2000: the 18th Conference on Computational Linguistics*, pages 947–953, Saarbrücken, Germany.

Jinchao Zhang, Peerachet Porkaew, Jiawei Hu, Qiuye Zhao, and Qun Liu. 2017. CASICT-DCU Neural Machine Translation Systems for WMT17. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark. Association for Computational Linguistics.

Yuan Zhang and David Weiss. 2016. Stack-propagation: Improved representation learning for syntax. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1557–1566, Berlin, Germany. Association for Computational Linguistics.