Findings of the 2016 Conference on Machine Translation (WMT16)

Ondřej Bojar, Charles University
Rajen Chatterjee, FBK
Christian Federmann, Microsoft Research
Yvette Graham, Dublin City University
Barry Haddow, Univ. of Edinburgh
Matthias Huck, LMU Munich
Antonio Jimeno Yepes, IBM Research Australia
Philipp Koehn, JHU / Edinburgh
Varvara Logacheva, Univ. of Sheffield
Christof Monz, Univ. of Amsterdam
Matteo Negri, FBK
Aurélie Névéol, LIMSI-CNRS
Mariana Neves, HPI/Potsdam
Martin Popel, Charles University
Matt Post, Johns Hopkins Univ.
Raphael Rubino, Saarland University
Carolina Scarton, Univ. of Sheffield
Lucia Specia, Univ. of Sheffield
Marco Turchi, FBK
Karin Verspoor, Univ. of Melbourne
Marcos Zampieri, Saarland University

Abstract

This paper presents the results of the WMT16 shared tasks, which included five machine translation (MT) tasks (standard news, IT-domain, biomedical, multimodal, pronoun), three evaluation tasks (metrics, tuning, run-time estimation of MT quality), and an automatic post-editing task and bilingual document alignment task. This year, 102 MT systems from 24 institutions (plus 36 anonymized online systems) were submitted to the 12 translation directions in the news translation task. The IT-domain task received 31 submissions from 12 institutions in 7 directions and the Biomedical task received 15 submissions systems from 5 institutions. Evaluation was both automatic and manual (relative ranking and 100-point scale assessments).

The quality estimation task had three subtasks, with a total of 14 teams, submitting 39 entries. The automatic post-editing task had a total of 6 teams, submitting 11 entries.

1 Introduction

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

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

- news translation (§2, §3)
- IT-domain translation (§4)
- biomedical translation (§5)
- quality estimation (§6)
- automatic post-editing (§7)

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

- tuning (Jawaid et al., 2016)
- metrics (Bojar et al., 2016b)
- cross-lingual pronoun prediction (Guillou et al., 2016)
- multimodal machine translation and crosslingual image description (Specia et al., 2016)
- bilingual document alignment (Buck and Koehn, 2016)

In the news translation task (§2), participants were asked to translate a shared test set, optionally restricting themselves to the provided training data. We held 12 translation tasks this year, between English and each of Czech, German, Finnish, Russian, Romanian, and Turkish. The Romanian and Turkish translation tasks were new this year, providing a lesser resourced data condition on challenging language pairs. The system outputs for each task were evaluated both automatically and manually.

The human evaluation (§3) involves asking human judges to rank sentences output by anonymized systems. We obtained large numbers of rankings from researchers who contributed

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 131–198,
Berlin, Germany, August 11-12, 2016.

evaluations proportional to the number of tasks they entered. We made data collection more efficient and used TrueSkill as ranking method. We also explored a novel way of ranking machine translation systems by judgments of adequacy and fluency on a 100-point scale.

The IT translation task (§4) was introduced this year and focused on domain adaptation of MT to the IT (information technology) domain and translation of answers in a cross-lingual help-desk service, where hardware&software troubleshooting answers are translated from English to the users' languages: Bulgarian, Czech, German, Spanish, Basque, Dutch and Portuguese. Similarly as in the News translation task, training and test data were provided and the system outputs were evaluated both automatically and manually.

Another task newly introduced this year was the biomedical translation task (§5). Participants were asked to translate the titles and abstracts of scientific articles indexed in the Scielo database. Training and test data were provided for two subdomains, biological sciences and health sciences, and three language pairs, Portuguese/English, Spanish/English and French/English. This task therefore provided data for a language not previously covered in WMT, Portuguese. The system outputs for each language pair were evaluated both automatically and manually.

The quality estimation task (§6) this year included three subtasks: sentence-level prediction of post-editing effort scores, word and phrase-level prediction of good/bad labels, and document-level prediction of human post-editing scores. Datasets were released with English→German IT translations for sentence and word/phrase level, and English↔Spanish news translations for document level.

The automatic post-editing task (§7) 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 second round, the task focused on correcting English→German translations in the IT 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.[1] We hope these datasets serve as a valuable resource for research into statistical machine translation and automatic evaluation or prediction of translation quality. News and IT 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, and Finnish. New languages this years are Romanian and Turkish.

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, Romanian), by Yandex[2], 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). For Finnish, a second translation was provided as well, but not used in the evaluation. All of the translations were done directly, and not via an intermediate language.

For Turkish we also released an additional 500 sentence development set, and for Romanian a third of the test set were released as a development

[1] `http://statmt.org/wmt16/results.html`
[2] `http://www.yandex.com/`

set instead. For the other languages, test sets from previous years are available as development sets.

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[3], United Nations, French-English 10^9 corpus, Common Crawl, Russian-English parallel data provided by Yandex, Wikipedia Headlines provided by CMU) and some were updated (CzEng v1.6pre (Bojar et al., 2016a), News Commentary v11, monolingual news data).

We added a few new corpora:

- Romanian Europarl (Koehn, 2002)
- SETIMES2 from OPUS for Romanian–English and Turkish–English (Tiedemann, 2009)
- Monolingual data sets from CommonCrawl (Buck et al., 2014)

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

2.3 Submitted systems

We received 102 submissions from 24 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 included 36 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

Each year, we conduct a human evaluation campaign to assess translation quality and determine the final ranking of candidate systems. This section describes how we prepared the evaluation data, collected human assessments, and computed the official results.

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

Over the past few years, our method of collecting and evaluating the manual translations has settled into the following pattern. We ask human annotators to rank the outputs of five systems. From these rankings, we produce pairwise translation comparisons, and then evaluate them with a version of the TrueSkill algorithm adapted to our task. We refer to this approach (described in Section 3.4) as the *relative ranking* approach (RR), so named because the pairwise comparisons denote only relative ability between a pair of systems, and cannot be used to infer their absolute quality. These results are used to produce the official ranking for the WMT 2016 tasks. However, work in evaluation over the past few years has provided fresh insight into ways to collect *direct assessments* (DA) of machine translation quality. In this setting, annotators are asked to provide an assessment of the direct quality of the output of a system relative to a reference translation. In order to evaluate the potential of this approach for future WMT evaluations, we conducted a direct assessment evaluation in parallel. This evaluation, together with a comparison of the official results, is described in Section 3.5.

3.1 Evaluation campaign overview

Following the trend from previous years, WMT16 ended up being the largest evaluation campaign to date. Similar to last year, we collected *researcher-based judgments* only (as opposed to crowdsourcing annotations from a tool like Mechanical Turk). For the News translation task, a total of 150 individual annotator accounts were involved. Users came from 33 different research groups and contributed judgments on 10,833 HITs.

Each HIT comprises three 5-way ranking tasks for a total of 32,499 such tasks. Under ordinary circumstances, each of the tasks would correspond to ten individual pairwise system comparisons denoting whether a system A was judged better than, worse than, or equivalent to another system B. However, since many systems have produced the same outputs for a particular sentence, we are often able to produce more than ten comparisons (Section 3.2), ending up with a total of 569,287 pairwise annotations—a 75.2% increase over the expected baseline of 324,990 pairs. This is smaller than last year's gain of 87.1% as we have decided to preserve punctuation differences. Section 3.2 provides more details on our pre-processing.

Europarl Parallel Corpus

	German ↔ English		Czech ↔ English		Finnish ↔ English		Romanian ↔ English	
Sentences	1,920,209		646,605		1,926,114		399,375	
Words	50,486,398	53,008,851	14,946,399	17,376,433	37,814,266	52,723,296	10,943,404	10,891,847
Distinct words	381,583	115,966	172,461	63,039	693,963	115,896	73,353	42,650

News Commentary Parallel Corpus

	German ↔ English		Czech ↔ English		Russian ↔ English	
Sentences	242,770		191,432		174,253	
Words	6,284,116	6,307,244	4,385,588	4,914,094	4,452,010	4,681,362
Distinct words	153,835	68,039	154,044	62,043	151,228	55,382

Common Crawl Parallel Corpus

	French ↔ English		German ↔ English		Czech ↔ English		Russian ↔ English	
Sentences	3,244,152		2,399,123		161,838		878,386	
Words	91,328,790	81,096,306	54,575,405	58,870,638	3,529,783	3,927,378	21,018,793	21,535,122
Distinct words	889,291	859,017	1,640,835	823,480	210,170	128,212	764,203	432,062

United Nations Parallel Corpus

	French ↔ English	
Sentences	12,886,831	
Words	411,916,781	360,341,450
Distinct words	565,553	666,077

Yandex 1M Parallel Corpus

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

10^9 Word Parallel Corpus

	French ↔ English	
Sentences	22,520,400	
Words	811,203,407	668,412,817
Distinct words	2,738,882	2,861,836

CzEng Parallel Corpus

	Czech ↔ English	
Sentences	51,424,584	
Words	592,890,104	699,087,647
Distinct words	3,073,115	1,727,574

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 words	282,989	251,328	127,576	96,732

Europarl Language Model Data

	English	German	Czech	Finnish
Sentence	2,218,201	2,176,537	668,595	2,120,739
Words	59,848,044	53,534,167	14,946,399	39,511,068
Distinct words	123,059	394,781	172,461	711,868

News Language Model Data

	English	German	Czech	Russian	Finnish	Romanian
Sentence	145,573,876	187,008,695	53,383,346	56,371,276	6,740,879	2,280,642
Words	3,355,935,396	3,331,396,767	879,993,532	1,016,368,612	83,112,454	54,793,949
Distinct words	5,487,137	16,166,174	3,824,351	3,834,224	2,572,117	504,438

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

	German ↔ EN		Czech ↔ EN		Russian ↔ EN		Finnish ↔ EN		Romanian ↔ EN		Turkish ↔ EN	
Sent.	2,999		2,999		2,998		3,000		1,999		2,998	
Words	64,379	65,647	57,097	66,457	62,840	71,068	48,839	64,611	50,603	48,531	54,420	67,468
Dist.	12,234	8,877	15,163	8,639	16,304	8,963	16,092	8,413	9,851	6,953	15,395	8,799

Figure 1: 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.

Language	Sources (Number of Documents)
English	ABC News (5), BBC (5), Brisbane Times (2), CBS News (2), CNN (1), Christian Science Monitor (2), Daily Mail (4), Euronews (1), Fox News (2), Guardian (9), Independent (1), Los Angeles Times (3), Medical Daily (1), News.com Australia (4), New York Times (1), Reuters (3), Russia Today (2), Scotsman (2), Sky (1), Sydney Morning Herald (5), stv.tv (1), Telegraph (4), The Local (2), Time Magazine (1), UPI (3), Xinhua Net (1).
Czech	aktuálně.cz (2), blesk.cz (3), deník.cz (8), e15.cz (2), iDNES.cz (12), ihned.cz (4), lidovky.cz (7), Novinky.cz (1), tyden.cz (6), ZDN (1).
German	Wirtschaftsblatt (1), Abendzeitung München (1), Abendzeitung Nürnberg (1), Ärztezeitung (1), Aachener Nachrichten (4), Berliner Kurier (1), Borkener Zeitung (1), Come On (1), Die Presse (2), Dülmener Zeitung (2), Euronews (1), Frankfurter Rundschau (1), Göttinger Tageblatt (1), Hessische/Niedersächsische Allgemeine (1), In Franken (4), Kleine Zeitung (3), Kreisanzeiger (1), Kreiszeitung (1), Krone (2), Lampertheimer Zeitung (1), Lausitzer Rundschau (1), Merkur (2), Morgenweb (1), Mitteldeutsche Zeitung (1), NTV (2), Nachrichten.at (6), Neues Deutschland (2), Neue Presse Coburg (1), Neue Westfälische (1), Ostfriesenzeitung (2), Passauer Neue Presse (1), Rheinzeitung (1), Schwarzwälder Bote (1), Segeberger Zeitung (1), Stuttgarter Nachrichten (1), Südkurier (3), Tagesspiegel (1), Teckbote (1), Thueringer Allgemeine (1), Thüringische Landeszeitung (1), tz München (1), Usinger Anzeiger (6), Volksblatt (3), Westfälischer Anzeiger (1), Weser Kurier (1), Wiesbadener Kurier (2), Westfälische Nachrichten (4), Westdeutsche Zeitung (3), Willhelmshavener Zeitung (1), Yahoo (1).
Finnish	Aamulehti (4), Etelä-Saimaa (2), Etelä-Suomen Sanomat (1), Helsingin Sanomat (12), Ilkka (5), Iltalehti (10), Ilta-Sanomat (31), Kaleva (3), Karjalainen (7), Kouvolan Sanomat (2).
Russian	168.ru (1), aif (2), altapress.ru (2), argumenti.ru (1), BBC Russian (1), Euronews (2), Fakty (3), Russia Today (1), Izvestiya (3), Kommersant (13), Lenta (7), lgng (2), MK RU (1), New Look Media (1), Novaya Gazeta (3), Novinite (1), ogirk.ru (1), pnp.ru (2), rg.ru (1), Rosbalt (2), rusplit.ru (1), Sport Express (10), trud.ru (2), tumentoday.ru (1), Vedomosti (1), Versia (2), Vesti (11), VM News (1).
Romanian	National (1), HotNews (1), Info Press (1), Puterea (1), ziare.ro (29), Ziarul de Iaşi (17)
Turkish	hurriyet (37), Sabah (26), Zaman (23)

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.

In total, our human annotators spent nearly 39 days and 3 hours working in Appraise. This gives an average annotation time of 6.4 hours per user. The average annotation time per HIT amounts to 5 minutes and 12 seconds. This is a little slower than last year's average time of 4 minutes and 53 seconds. Similar to the previous campaign, several of the annotators passed the mark of more than 100 HITs annotated (the maximum number being 684) and, again, some worked for more than 24 hours (the most patient annotator contributing a little over 99 hours of annotation work).

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

3.2 Data collection

The system ranking is produced from a large set of pairwise judgments, each of which indicates the relative quality of the outputs of two systems' translations of the same input sentence. Annotations are collected in an evaluation campaign that enlists participants in the shared task to help. Each team is asked to contribute one hundred so-called "Human Intelligence Tasks" (HITs) per primary system submitted.

We continue to use the open-source Appraise[4] (Federmann, 2012) tool for our data collection. Last year, we had provided the following instructions at the top of each HIT page:

> *You are shown a source sentence followed by several candidate translations. Your task is to rank the translations from best to worst (ties are allowed).*

This year, in order to optimize screen space we have streamlined the user interface, removing the instruction text (which instead was communicated to annotators outside of the HIT annotation interface) and trimming vertical spacing. A screenshot of the Appraise relative ranking interface is shown in Figure 2.

Annotators are asked to rank the outputs from 1 (best) to 5 (worst), with ties permitted. Note that a *lower* rank is better, and that this is clear from the interface design. Annotators can decide to skip a ranking task but are instructed to do this only as a last resort, e.g., if the translation candidates shown on screen are clearly misformatted or contain data

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

ID	Institution
AALTO	Aalto University (Grönroos et al., 2016)
ABUMATRAN-*	Abu-MaTran (Sánchez-Cartagena and Toral, 2016)
AFRL-MITLL	Air Force Research Laboratory / MIT Lincoln Lab (Gwinnup et al., 2016)
AMU-UEDIN	Adam Mickiewicz Uni. / Uni. Edinburgh (Junczys-Dowmunt et al., 2016)
CAMBRIDGE	University of Cambridge (Stahlberg et al., 2016)
CMU	Carnegie Mellon University
CU-MERGEDTREES	Charles University (Mareček, 2016)
CU-CHIMERA CU-TAMCHYNA	Charles University (Tamchyna et al., 2016)
CU-TECTOMT	Charles University (Dušek et al., 2015)
JHU-*	Johns Hopkins University (Ding et al., 2016)
KIT, KIT-LIMSI	Karlsruhe Institute of Technology (Ha et al., 2016)
LIMSI	University of Paris (Allauzen et al., 2016)
LMU-CUNI	University of Munich / Charles University (Tamchyna et al., 2016)
METAMIND	Salesforce Metamind (Bradbury and Socher, 2016)
NRC	National Research Council Canada (Lo et al., 2016)
NYU-MONTERAL	New York University / University of Montréal (Chung et al., 2016)
PARFDA	Ergun Bicici (Bicici, 2016a)
PJATK	Polish-Japanese Academy of Inf. Technology (Wołk and Marasek, 2016)
PROMT	PROMT Automated Translation Solutions (Molchanov and Bykov, 2016)
QT21-HIML	QT21 System Combination (Peter et al., 2016b)
RWTH	RWTH Aachen (Peter et al., 2016a)
TBTK	TÜBITAK (Bektaş et al., 2016)
UEDIN-NMT	University of Edinburgh (Sennrich et al., 2016)
UEDIN-PBMT UEDIN-SYNTAX	University of Edinburgh (Williams et al., 2016)
UEDIN-LMU	University of Edinburgh / University of Munich (Huck et al., 2016)
UH-*	University of Helsinki (Tiedemann et al., 2016)
USFD-RESCORING	University of Sheffield (Blain et al., 2016)
UUT	Uppsala University (Tiedemann et al., 2016)
YSDA	Yandex School of Data Analysis (Dvorkovich et al., 2016)
ONLINE-[A,B,F,G]	Four online statistical machine translation systems

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.

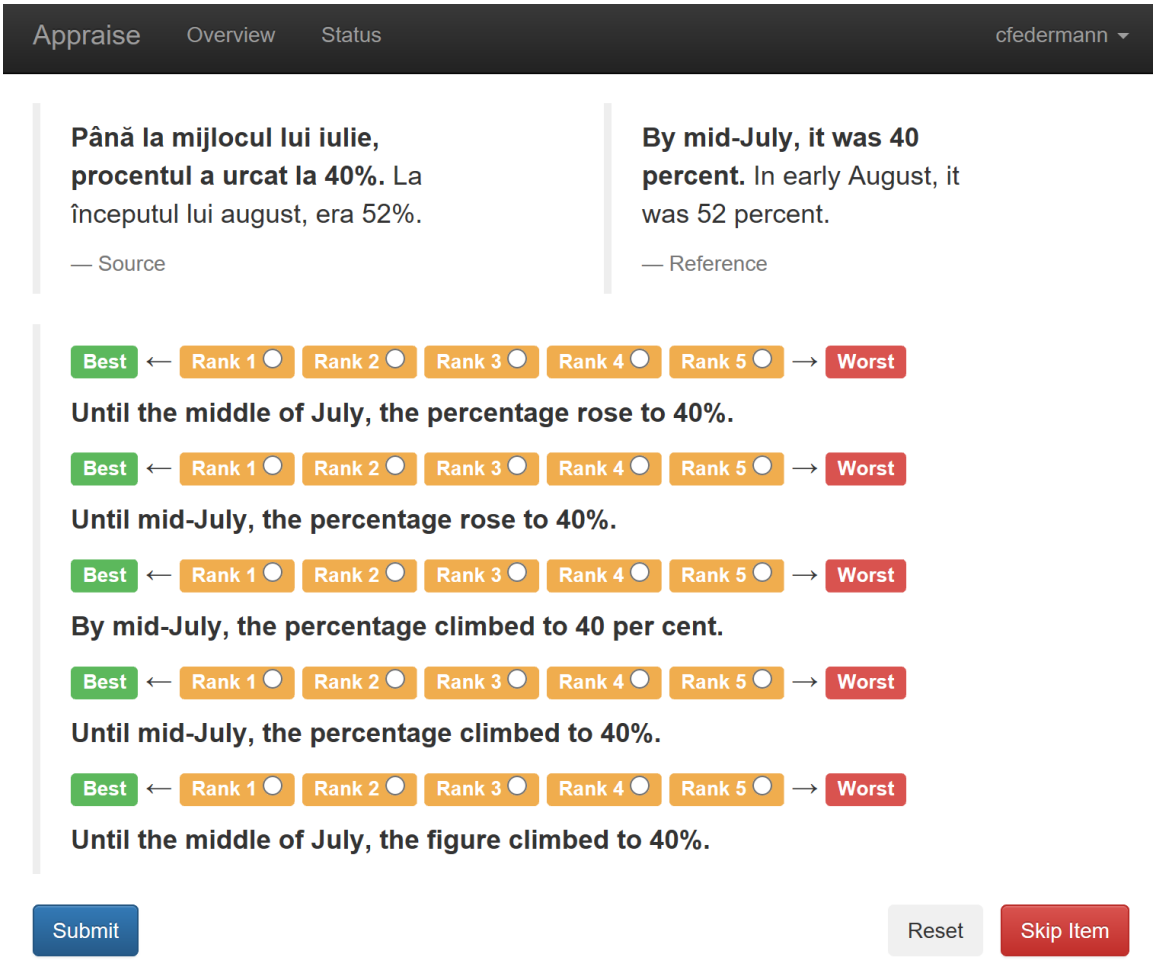

Figure 2: Screenshot of the Appraise interface used in the human evaluation campaign. The annotator is presented with a source segment, a reference translation, and up to five outputs from competing systems (anonymized and displayed in random order), and is asked to rank these according to their translation quality, with ties allowed.

issues (wrong language, encoding errors or other, obvious problems). Similar to last year, only a few ranking tasks have been skipped in WMT16.

Each HIT consists of three so-called *ranking tasks*. In a ranking task, an annotator is presented with a source segment, a human reference translation, and the outputs of up to five anonymized candidate systems, randomly selected from the set of participating systems, and displayed in random order. This year, as with last year, we perform a redundancy cleanup as an initial preprocessing step and create *multi-system outputs* to avoid confusing annotators with identical content: instead of selecting five systems and displaying their (identical) outputs, we select five *distinct* outputs, and then propagate the collected rankings to all the individual systems within each of the respective multi-system outputs. Last year, however, nearly-identical outputs were collapsed if they differed only on punctuation. Because punctuation is an important component of producing quality MT output, this year, we only collapse outputs that are exactly the same, apart from differences in nonzero whitespace.

To demonstrate how this works, we provide the following example. First, consider the case where we select system outputs directly, instead of the multi-system outputs described above. Here, we consider an annotation provided by a judge among the outputs of systems A, B, F, H, and J:

	1	2	3	4	5
F				•	
A				•	
B		•			
J					•
H			•		

The joint rankings provided by a ranking task are then expanded to a set of *pairwise rankings* produced by considering all $\binom{n}{2} \leq 10$ combinations of all $n \leq 5$ outputs in the respective ranking task.

Language Pair	Systems	Comparisons	Comparisons/Sys
Czech→English	12	125,788	10,482.3
English→Czech	20	192,487	9,624.3
Finnish→English	9	30,519	3,391.0
English→Finnish	13	38,254	2,942.6
German→English	10	20,937	2,093.7
English→German	15	50,989	3,399.2
Romanian→English	7	15,822	2,260.2
English→Romanian	12	11,352	946.0
Russian→English	10	27,353	2,735.3
English→Russian	12	34,414	2,867.8
Turkish→English	9	10,188	1,132.0
English→Turkish	9	11,184	1,242.6
Totals WMT16	138	569,287	4,125.2
WMT15	131	542,732	4,143.0
WMT14	110	328,830	2,989.3
WMT13	148	942,840	6,370.5
WMT12	103	101,969	999.6
WMT11	133	63,045	474.0

Table 3: Amount of data (pairwise comparisons after "de-collapsing" *multi-system outputs*) collected in the WMT16 manual evaluation campaign. The final five 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 tuning shared task.

As the number of outputs n depends on the number of identical (and, hence, redundant) *multi-system outputs* in the original data, we end up getting varying numbers of corresponding binary judgments. Now, consider the case of *multi-system outputs*. If the outputs of system A and F from above are actually identical, the annotator this year would see an easier ranking task:[5]

	1	2	3	4	5
AF				•	
B		•			
J					•
H			•		

Both examples would be reduced to the following set of pairwise judgments:

$$A > B, A = F, A > H, A < J$$
$$B < F, B < H, B < J$$
$$F > H, F < J$$
$$H < J$$

Here, $A > B$ should be read is "A is ranked higher than (worse than) B". Note that by this procedure, the absolute value of ranks and the magnitude of their differences are discarded. In the case of multi-system outputs, this set of pairwise rankings would have been produced with less annotator effort. This productivity gain grows in the number of systems that produce identical output, and this situation is quite common, due in part to the fact that many systems are built on the same underlying technology. Table 3 has more details.

[5]Technically, another distinct output would have been inserted, if possible, so as to present the annotator with five, but we ignore that for illustration purposes.

3.3 Annotator agreement

Each year we calculate annotator agreement scores for the human evaluation as a measure of the reliability of the rankings. We measured pairwise agreement among annotators using Cohen's kappa coefficient (κ) (Cohen, 1960). If $P(A)$ be the proportion of times that the annotators agree, and $P(E)$ is the proportion of time that they would agree by chance, then Cohen's kappa is:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

Note that κ is basically a normalized version of $P(A)$, one which takes into account how meaningful it is for annotators to agree with each other by incorporating $P(E)$. The values for κ range from 0 to 1, with zero indicating no agreement and 1 perfect agreement.

We calculate $P(A)$ by examining all pairs of

Language Pair	WMT12	WMT13	WMT14	WMT15	WMT16
Czech→English	0.311	0.244	0.305	0.458	0.244
English→Czech	0.359	0.168	0.360	0.438	0.381
German→English	0.385	0.299	0.368	0.423	0.475
English→German	0.356	0.267	0.427	0.423	0.369
French→English	0.272	0.275	0.357	0.343	—
English→French	0.296	0.231	0.302	0.317	—
Russian→English	—	0.278	0.324	0.372	0.339
English→Russian	—	0.243	0.418	0.336	0.340
Finnish→English	—	—	—	0.388	0.293
English→Finnish	—	—	—	0.549	0.484
Romanian→English	—	—	—	—	0.379
English→Romanian	—	—	—	—	0.341
Turkish→English	—	—	—	—	0.322
English→Turkish	—	—	—	—	0.319
Mean	0.330	0.260	0.367	0.405	0.357

Table 4: κ scores measuring inter-annotator agreement for WMT16. See Table 5 for corresponding intra-annotator agreement scores. WMT14–WMT16 results are based on researchers' judgments only, whereas prior years mixed judgments of researchers and crowdsourcers.

outputs[6] which had been judged by two or more judges, and calculating the proportion of time that they agreed that $A < B$, $A = B$, or $A > B$. In other words, $P(A)$ is the empirical, observed rate at which annotators agree, in the context of pairwise comparisons.

As for $P(E)$, it captures the probability that two annotators would agree randomly. Therefore:

$$P(E) = P(A<B)^2 + P(A=B)^2 + P(A>B)^2$$

Note that each of the three probabilities in $P(E)$'s definition are squared to reflect the fact that we are considering the chance that *two* annotators would agree by chance. Each of these probabilities is computed empirically, by observing how often annotators actually rank two systems as being tied.

Table 4 shows final κ values for inter-annotator agreement for WMT11–WMT16 while Table 5 details intra-annotator agreement scores. The exact interpretation of the kappa coefficient is difficult, but according to Landis and Koch (1977), 0–0.2 is *slight*, 0.2–0.4 is *fair*, 0.4–0.6 is *moderate*, 0.6–0.8 is *substantial*, and 0.8–1.0 is *almost perfect*.

Compared to last year's results, inter-annotator agreement rates have decreased. Notably, for Czech→English, we see a drop from 0.458 to 0.244. English→Czech decreases from 0.438 to 0.381. Considering that the total number of data points collected as well as the number of annotators for these language pairs have increased substantially, the lower agreement score seems plausible.[7] We observe a small increase in agreement for German→English (from 0.423 to 0.475) and a drop for English→German (from 0.434 to 0.369). Scores for both Russian language pairs are similar to what had been measured in WMT15. For Finnish, we again see a decrease (from 0.388 to 0.293 for Finnish→English and from 0.549 to 0.484 for English→Finnish) and our new languages, Romanian and Turkish, end up with *fair* annotator agreement. The average inter-annotator agreement across all languages is 0.357, which is also *fair* and comparable to researchers' agreement over the last years. Intra-annotator agreement scores have mostly decreased compared to WMT15, except for both Russian language pairs. The new languages show *moderate* agreement except for English→Turkish which achieves a *fair* score. On average we observe an intra-annotator agreement which is comparable to researcher-based scores from WMT13–WMT15.

[6]Regardless if they correspond to an individual system or to a set of systems ("multi-system") producing identical translations. Thus, when computing annotator agreement scores, we effectively treat both individual and multi-systems in the same way, as "individual comparison units". By doing so, we avoid artificially inflating our agreement scores based on the automatically inferred $A = B$ ties from multi-systems.

[7]Both Czech→English and English→Czech contain tuning-task systems with very similar quality (according to both human evaluation and BLEU), which makes the annotation task more difficult.

Language Pair	WMT12	WMT13	WMT14	WMT15	WMT16
Czech→English	0.454	0.479	0.382	0.694	0.504
English→Czech	0.390	0.290	0.448	0.584	0.438
German→English	0.392	0.535	0.344	0.801	0.552
English→German	0.433	0.498	0.576	0.676	0.529
French→English	0.360	0.578	0.629	0.510	—
English→French	0.414	0.495	0.507	0.426	—
Russian→English	—	0.450	0.629	0.506	0.552
English→Russian	—	0.513	0.570	0.492	0.528
Finnish→English	—	—	—	0.562	0.549
English→Finnish	—	—	—	0.697	0.617
Romanian→English	—	—	—	—	0.621
English→Romanian	—	—	—	—	0.552
Turkish→English	—	—	—	—	0.559
English→Turkish	—	—	—	—	0.352
Mean	0.407	0.479	0.522	0.595	0.529

Table 5: κ scores measuring intra-annotator agreement, i.e., self-consistency of judges, across for the past few years of the human evaluation campaign. Scores are in line with results from WMT14 and WMT15.

3.4 Producing the human ranking

The collected pairwise rankings are used to produce the official human ranking of the systems. Since WMT14, we have used the TrueSkill method for producing the official ranking, in the following fashion. We produce 1,000 bootstrap-resampled datasets over all of the available data (i.e., datasets sampled uniformly with replacement from the complete dataset). We run TrueSkill over each dataset. We then compute a *rank range* for each system by collecting the absolute rank of each system in each fold, throwing out the top and bottom 2.5%, and then clustering systems into equivalence classes containing systems with overlapping ranges, yielding a partial ordering over systems at the 95% confidence level.

The full list of the official human rankings for each task can be found in Table 6, which also reports all system scores, rank ranges, and clusters for all language pairs and all systems. The official interpretation of these results is that systems in the same cluster are considered tied. Given the large number of judgments that we collected, it was possible to group on average about two systems in a cluster, even though the systems in the middle are typically in larger clusters.

In Figure 3–5, we plotted the human evaluation result against everybody's favorite metric BLEU. Although these two metrics correlate generally well, the plots clearly suggest that a fair comparison of systems of different kinds cannot rely on automatic scores. Rule-based systems receive a much lower BLEU score than statistical systems (see for instance English–German, e.g., PROMT-RULE). The same is true to a lesser degree for statistical syntax-based systems (see English–German, UEDIN-SYNTAX vs. UEDIN-PBMT).

3.5 Direct Assessment Manual Evaluation

In addition to the standard relative ranking (RR) manual evaluation, this year a new method of human evaluation was also trialed in the main translation task: monolingual direct assessment (DA) of translation fluency (Graham et al., 2013) and adequacy (Graham et al., 2014, 2016).

Agreement between human assessors of translation quality is a known problem in evaluation of MT and DA therefore aims to simplify translation assessment, which conventionally takes the form of a bilingual evaluation, by restructuring the task into a monolingual assessment. Figure 6 provides a screen shot of DA adequacy assessment, where the task is structured as a monolingual similarity of meaning task.

Human assessors are asked to rate a given translation by how adequately it expresses the meaning of the corresponding reference translation on an analogue scale, which corresponds to an underlying absolute 0–100 rating. DA fluency assessment is similar with two exceptions, firstly no reference translation is displayed and secondly, assessors are asked to rate how much they agree that a given translation is fluent target language text. DA flu-

Czech–English

#	score	range	system
1	0.62	1	UEDIN-NMT
2	0.32	2	JHU-PBMT
3	0.21	3	ONLINE-B
4	0.11	4-6	TT-BLEU-MIRA
	0.10	4-7	TT-AFRL
	0.09	4-7	TT-NRC-NNBLEU
	0.07	5-8	TT-NRC-MEANT
	0.03	7-10	TT-BEER-PRO
	0.00	8-10	PJATK
	0.00	8-10	TT-BLEU-MERT
5	−0.07	11	ONLINE-A
6	−1.48	12	CU-MRGTREES

English–Czech

#	score	range	system
1	0.59	1	UEDIN-NMT
2	0.43	2	NYU-MONTREAL
3	0.34	3	JHU-PBMT
4	0.30	4-5	CU-CHIMERA
	0.30	4-5	CU-TAMCHYNA
5	0.22	6-7	UEDIN-CU-SYTX
	0.19	6-7	ONLINE-B
6	0.16	8-11	TT-BLEU-MIRA
	0.15	8-12	TT-BEER-PRO
	0.15	8-13	TT-BLEU-MERT
	0.14	9-14	TT-AFRL2
	0.14	9-14	TT-AFRL1
	0.13	9-14	TT-DCU
	0.13	11-14	TT-FJFI
7	0.08	15	ONLINE-A
8	−0.03	16	CU-TECTOMT
9	−0.43	17	TT-USAAR-HMM-MERT
10	−0.54	18	CU-MRGTREES
11	−1.13	19	TT-USAAR-HMM-MIRA
12	−1.33	20	TT-USAAR-HARM

Romanian–English

#	score	range	system
1	0.58	1-2	ONLINE-B
	0.38	1-2	UEDIN-NMT
2	0.10	3	UEDIN-PBMT
3	−0.09	4-5	UEDIN-SYNTAX
	−0.19	4-6	ONLINE-A
	−0.32	5-7	JHU-PBMT
	−0.46	6-7	LIMSI

English–Romanian

#	score	range	system
1	0.45	1-2	UEDIN-NMT
	0.43	1-2	QT21-HIML-COMB
2	0.20	3-7	KIT
	0.16	3-7	UEDIN-PBMT
	0.14	3-7	ONLINE-B
	0.14	3-7	UEDIN-LMU-HIERO
	0.12	3-7	RWTH-COMB
3	−0.15	8-10	LIMSI
	−0.23	8-10	LMU-CUNI
	−0.26	8-11	JHU-PBMT
	−0.43	10-12	USFD-RESCORING
	−0.57	11-12	ONLINE-A

German–English

#	score	range	system
1	0.82	1	UEDIN-NMT
2	0.25	2-5	ONLINE-B
	0.21	2-5	ONLINE-A
	0.19	2-5	UEDIN-SYNTAX
	0.18	2-6	KIT
	0.04	5-7	UEDIN-PBMT
	0.03	6-7	JHU-PBMT
3	−0.12	8	ONLINE-G
4	−0.67	9	JHU-SYNTAX
5	−0.93	10	ONLINE-F

Russian–English

#	score	range	system
1	0.45	1-2	AMU-UEDIN
	0.43	1-3	ONLINE-G
	0.33	2-4	NRC
	0.25	3-5	ONLINE-B
	0.16	4-5	UEDIN-NMT
2	0.04	6-7	ONLINE-A
	0.02	6-7	AFRL-MITLL-PHR
3	−0.11	8-9	AFRL-MITLL-CNTR
	−0.17	8-9	PROMT-RULE
4	−1.39	10	ONLINE-F

English–Russian

#	score	range	system
1	0.79	1	PROMT-RULE
2	0.30	2-4	AMU-UEDIN
	0.26	2-5	ONLINE-B
	0.26	2-5	UEDIN-NMT
	0.20	3-5	ONLINE-G
3	0.10	6	NYU-MONTREAL
4	−0.02	7-8	JHU-PBMT
	−0.07	7-10	LIMSI
	−0.10	8-10	ONLINE-A
	−0.15	9-10	AFRL-MITLL-PHR
5	−0.31	11	AFRL-MITLL-VERB
6	−1.26	12	ONLINE-F

Turkish–English

#	score	range	system
1	0.82	1-2	ONLINE-B
	0.65	1-3	ONLINE-G
	0.56	2-3	ONLINE-A
2	0.21	4-5	TBTK-SYSCOMB
	0.12	4-6	PROMT-SMT
	−0.00	5-6	YSDA
3	−0.67	7-8	JHU-SYNTAX
	−0.76	7-9	JHU-PBMT
	−0.94	8-9	PARFDA

English–German

#	score	range	system
1	0.49	1	UEDIN-NMT
2	0.40	2	METAMIND
3	0.29	3	UEDIN-SYNTAX
4	0.17	4	NYU-MONTREAL
5	−0.01	5-10	ONLINE-B
	−0.01	5-10	KIT-LIMSI
	−0.02	5-10	CAMBRIDGE
	−0.02	5-10	ONLINE-A
	−0.03	5-10	PROMT-RULE
	−0.05	6-10	KIT
6	−0.14	11-12	JHU-SYNTAX
	−0.15	11-12	JHU-PBMT
7	−0.26	13-14	UEDIN-PBMT
	−0.33	13-15	ONLINE-F
	−0.34	14-15	ONLINE-G

Finnish–English

#	score	range	system
1	0.42	1-4	UEDIN-PBMT
	0.40	1-4	ONLINE-G
	0.39	1-4	ONLINE-B
	0.34	1-4	UH-OPUS
2	0.01	5	PROMT-SMT
3	−0.11	6-7	UH-FACTORED
	−0.13	6-7	UEDIN-SYNTAX
4	−0.29	8	ONLINE-A
5	−1.03	9	JHU-PBMT

English–Finnish

#	score	range	system
1	0.36	1-3	ONLINE-G
	0.31	1-4	ABUMATRAN-NMT
	0.29	1-4	ONLINE-B
	0.23	3-5	ABUMATRAN-CMB
	0.16	4-5	UH-OPUS
2	−0.01	6-8	ABUMATRAN-PB
	−0.02	6-8	NYU-MONTREAL
	−0.02	6-8	ONLINE-A
3	−0.14	9-10	JHU-PBMT
	−0.23	9-12	UH-FACTORED
	−0.28	10-13	AALTO
	−0.30	10-13	JHU-HLTCOE
	−0.35	11-13	UUT

English–Turkish

#	score	range	system
1	0.76	1-2	ONLINE-G
	0.62	1-2	ONLINE-B
2	0.38	3	ONLINE-A
3	0.06	4	YSDA
4	−0.13	5-6	JHU-HLTCOE
	−0.19	5-7	TBTK-MORPH
	−0.29	6-7	CMU
5	−0.54	8-9	JHU-PBMT
	−0.66	8-9	PARFDA

Table 6: Official results for the WMT16 translation task. Systems are ordered by their inferred system means, though systems within a cluster are considered tied. Lines between systems indicate clusters according to bootstrap resampling at p-level $p \leq .05$. Systems with gray background indicate use of resources that fall outside the constraints provided for the shared task.

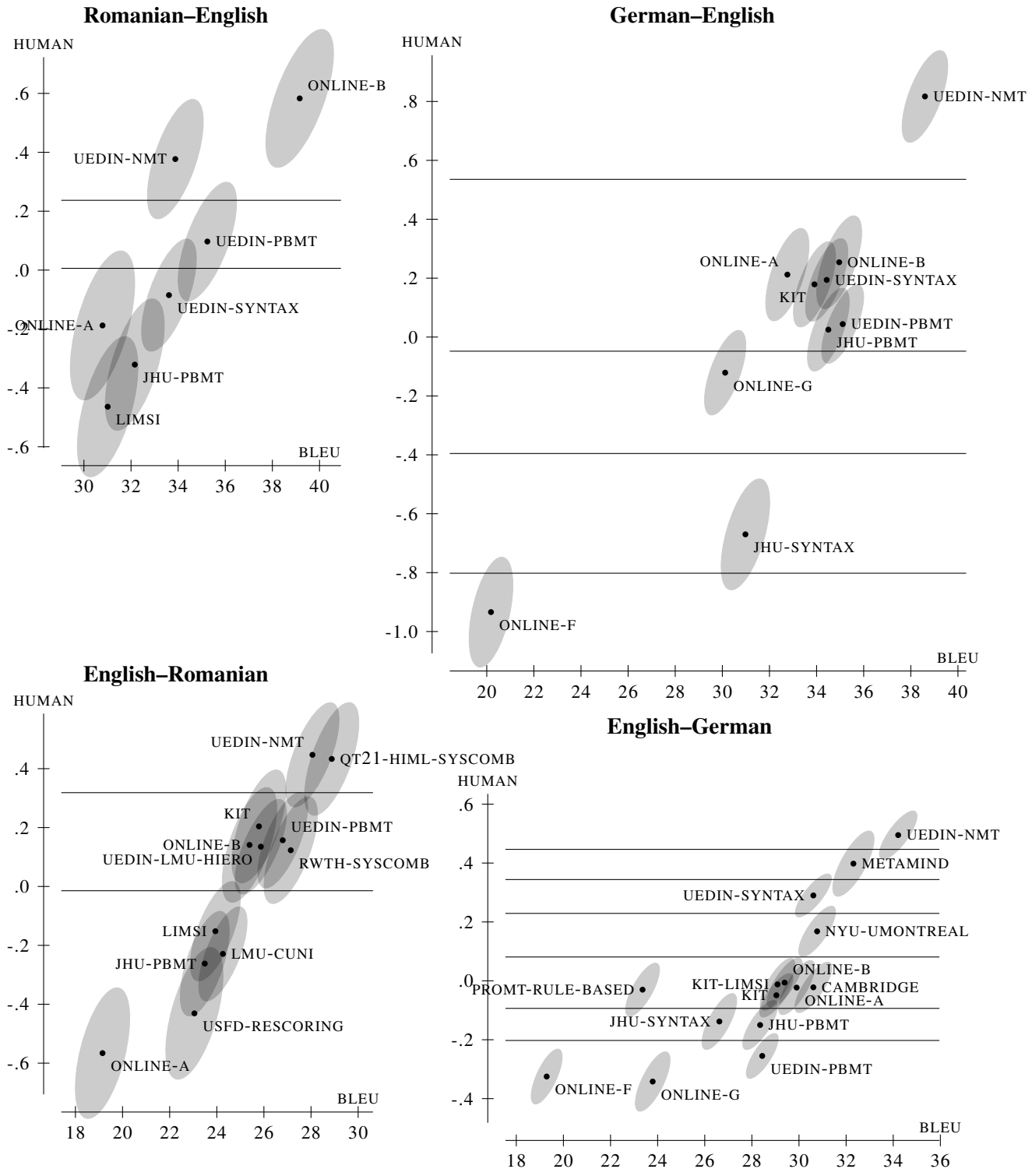

Figure 3: Human evaluation scores versus BLEU scores for the German–English and Romanian–English language pairs illustrate the need for human evaluation when comparing systems of different kind. Confidence intervals are indicated by the shaded ellipses. Rule-based systems and to a lesser degree syntax-based statistical systems receive a lower BLEU score than their human score would indicate. The big cluster in the Czech-English plot are tuning task submissions.

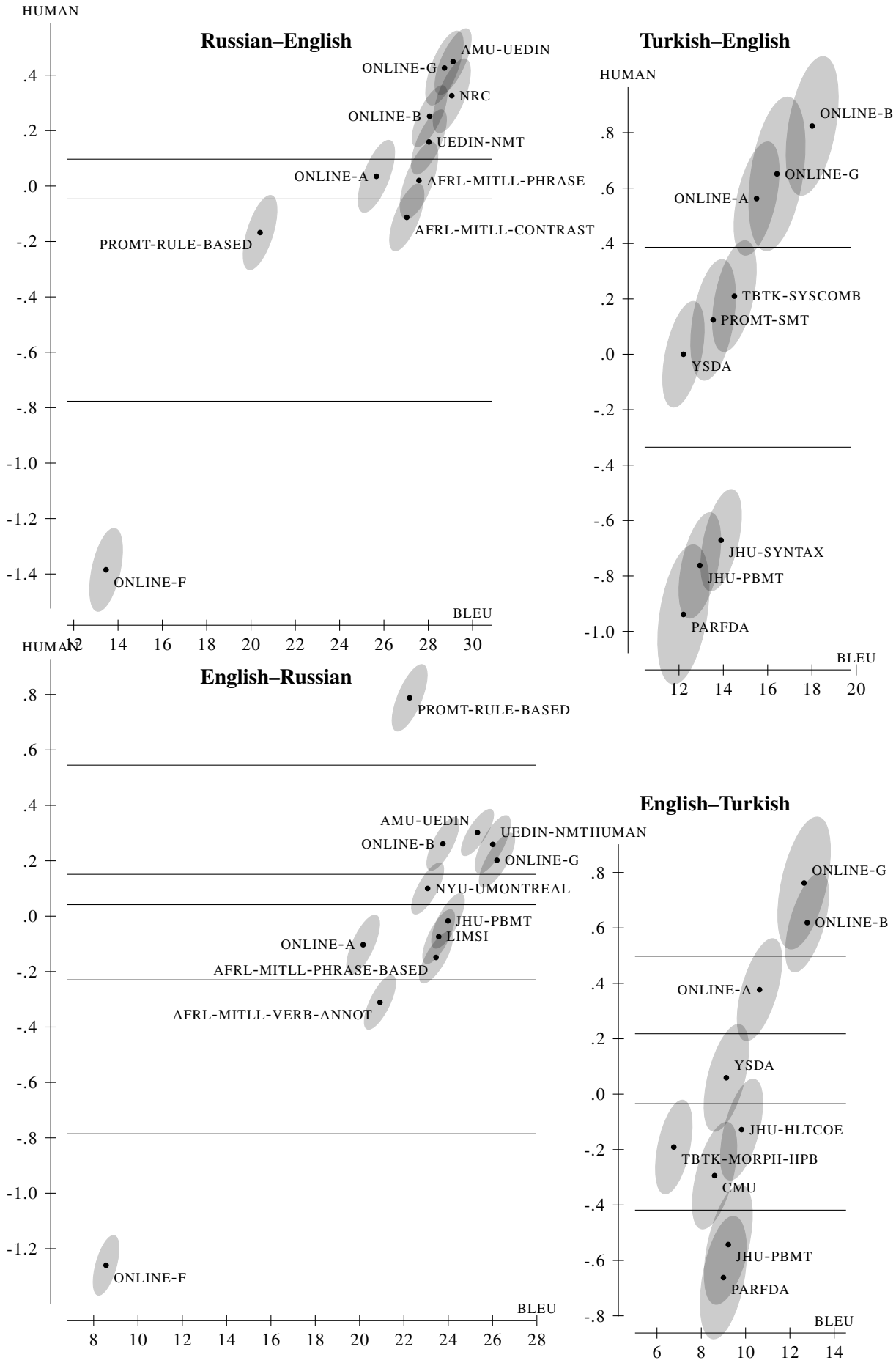

Figure 4: Human evaluation scores versus BLEU scores for the Russian–English and Turkish–English language pairs

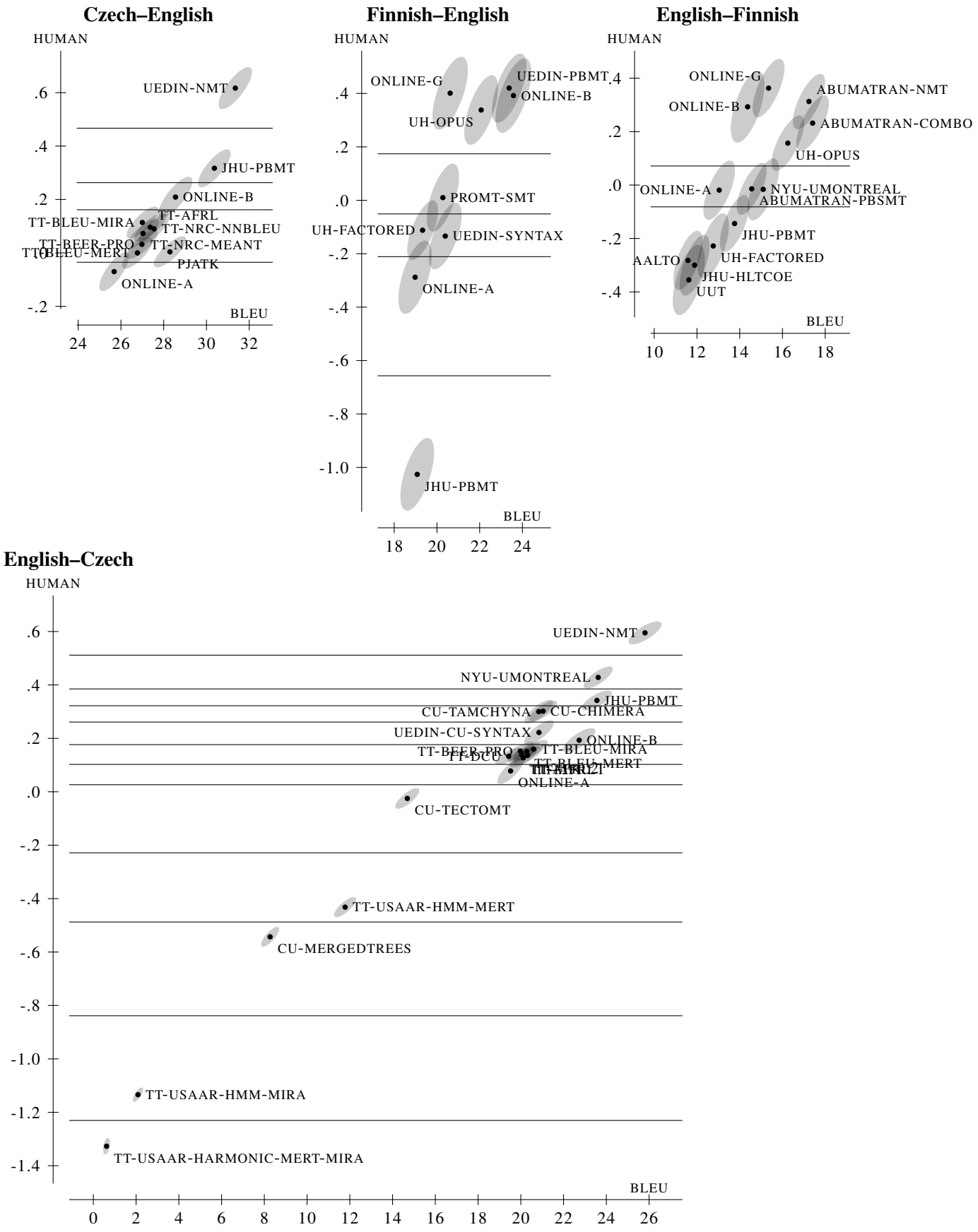

Figure 5: Human evaluation scores versus BLEU scores for the Czech–English and Finnish–English language pairs

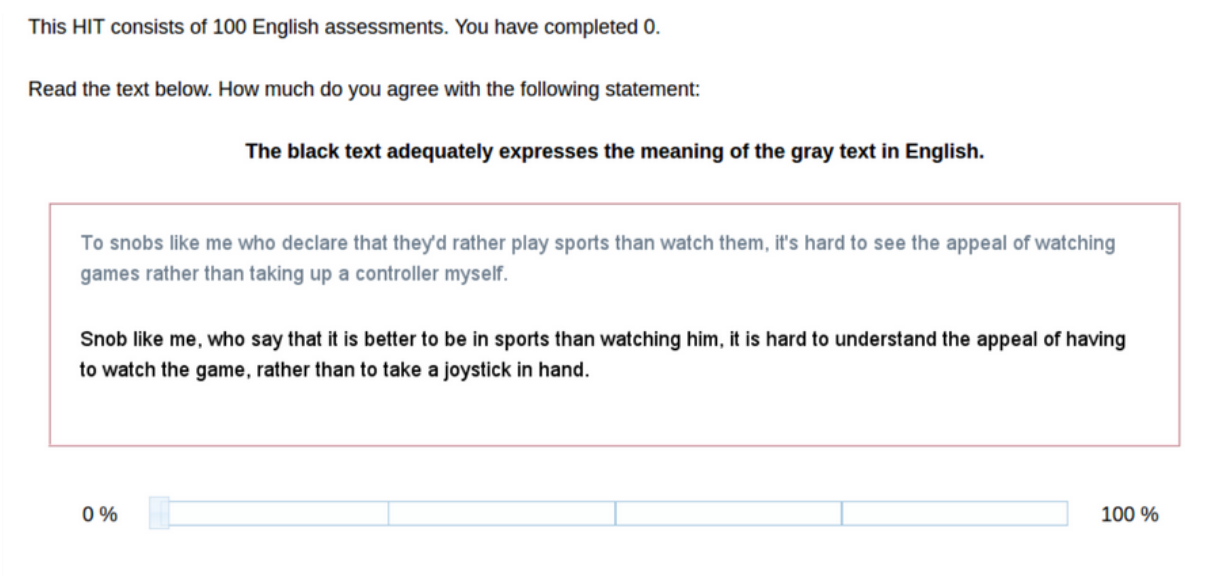

Figure 6: Direct Assessment of translation adequacy as carried out by workers on Mechanical Turk.

ency therefore provides a dimension of the assessment that cannot be biased by the presence of a reference translation. For both fluency and adequacy, the simpler monolingual assessment DA employs also allows the sentence length restriction to be removed.[8]

DA also aims to avoid the possible source of bias identified in Bojar et al. (2011), introduced by simultaneous assessment of several translations at once, where systems for which translations were more frequently compared to other low or high quality outputs resulted in either an unfair advantage or disadvantage for that system. We therefore elicit assessments of individual translations in isolation from the output of other systems, an important criteria when aiming for absolute quality judgments.

Large numbers of human assessments of translations for seven language pairs (cs-en, de-en, fi-en, ro-en, ru-en, tr-en and en-ru) were collected on Amazon's Mechanical Turk.[9] Table 7 shows overall numbers of translation assessments carried out.

Translations are arranged in sets of 100-translations per HIT to ensure sufficient repeat items per worker, before application of strict quality control measures to filter out assessments from poorly performing workers. When an analogue (or 100-points, 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 human assessors by how consistently they rate translations of known distinct quality.

A degraded version of a given original system output translation is automatically generated by substituting a sequence of words with a random phrase, itself selected from elsewhere in the reference document. Together with the original output, the degraded translation is known as a *bad reference* translation pair. Bad reference pairs are subsequently hidden within HITs, and provide a mechanism for filtering out workers who are simply not up to the task or those attempting to game the system. Assessments of workers who do not reliably score bad reference translations significantly lower than corresponding genuine system output translations are filtered out by comparison of scores they attribute to bad reference pairs within HITs. More specifically, we apply a paired Wilcoxon signed-rank test to score distributions of bad reference pairs, yielding a p-value for each worker we subsequently employ as a reliability estimate. Assessments of workers whose p-value lies above the conventional 0.05 threshold are omitted from the evaluation of systems.

Table 8 shows the number of unique workers who evaluated MT output on Mechanical Turk via DA for WMT16 for both fluency and adequacy, those who met our filtering requirement by show-

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

[9]www.mturk.com

	Adequacy			Fluency		
	Pre Quality Control	Post Quality Control	Ave. per System	Pre Quality Control	Post Quality Control	Ave. per System
cs-en	30,000	16,800 (56.0%)	2,800	16,880	6,880 (40.8%)	1,146
de-en	68,800	33,760 (49.1%)	3,376	20,480	10,400 (50.8%)	1,040
fi-en	63,040	30,080 (47.7%)	3,342	21,760	9,680 (44.5%)	1,075
ro-en	27,920	16,000 (57.3%)	2,285	18,960	8,000 (42.2%)	1,142
ru-en	64,960	37,040 (57.0%)	3,704	24,640	11,520 (46.8%)	1,152
tr-en	48,640	18,400 (37.8%)	2,044	28,000	10,640 (38.0%)	1,182
en-ru	38,160	15,920 (41.7%)	1,326	-	-	-
Overall	341,520	168,000 (49.2%)	2,666	130,720	57,120 (43.7%)	1,120

DA Manual Evaluation Assessments

Table 7: Numbers of system output translations evaluated on Mechanical Turk for direct assessment (DA) in WMT16, numbers exclude quality control items.

	All	(A) Sig. Diff. Bad Ref.	(A) & No Sig. Diff. Exact Rep.
Adequacy	1307	735	717 (98%)
Fluency	864	380	372 (98%)

DA Workers

Table 8: Number of unique human assessors for DA adequacy and fluency on Mechanical Turk in WMT16, (A) those whose scores for bad reference pairs were significantly different and numbers of unique human assessors in (A) whose scores for exact repeat items also showed no significant difference, paired Wilcoxon signed-rank significance test was applied in both cases.

ing 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 an identical previous translation.

In order to iron out differences in scoring strategies of distinct workers, human assessment scores for translations are standardized according to each individual worker's overall mean and standard deviation score. Subsequently, the overall score of a given MT system participating in the shared task simply comprises the mean (standardized) score of its translations.

Table 9 includes mean DA fluency and adequacy scores for all to-English systems participating in WMT16 translation task, while Table 10 includes results for the single out-of-English language pair for which DA was run this year, English to Russian. Mean standardized scores for systems not significantly lower than that of any other participating system, according to Wilcoxon signed-rank test, for a given language pair, are highlighted in bold. Although we also evaluated the fluency of translations, mean standardized adequacy scores should provide the primary mechanism for ranking competing systems, since it is entirely possible to achieve a high fluency score without conveying the meaning of the source input. Fluency can be employed as a secondary mechanism to break systems tied for adequacy or for diagnostic purposes. Figures 7, 8 and 9 show results of combining significance test conclusions for DA adequacy and fluency, where any ties between systems tied for adequacy are broken if that system outperformed the other with respect to fluency. It should be noted that RR provide official task results, while DA results are investigatory and do not indicate official translation task winners.

Finally, we compare scores of the official ranking to mean standardized adequacy scores for systems evaluated with DA. Table 11 shows the Pearson correlation between Trueskill scores for systems evaluated by researchers with relative preference judgments (official results) and DA mean scores collected via crowd-sourcing, showing high levels of agreement reached overall for all language pairs as correlations range from 0.92 to 0.997.

		DA Adequacy		**DA** Fluency	
		mean z	mean raw (%)	mean z	mean raw (%)
cs-en	UEDIN-NMT	**0.207**	75.4	**0.499**	78.7
	JHU-PBMT	0.101	72.6	0.194	69.3
	ONLINE-B	0.051	70.8	0.052	64.6
	ONLINE-A	0.000	69.5	−0.057	61.2
	PJATK	−0.024	69.0	−0.014	62.8
	CU-MERGEDTREES	−0.503	55.8	−0.754	41.1
de-en	UEDIN-NMT	**0.204**	75.8	**0.339**	77.5
	ONLINE-A	0.095	72.7	0.094	70.1
	ONLINE-B	0.086	72.2	0.015	68.4
	UEDIN-SYNTAX	0.065	71.5	0.141	71.8
	KIT	0.062	71.4	0.192	72.7
	UEDIN-PBMT	0.042	70.9	0.004	68.6
	JHU-PBMT	0.019	70.5	0.084	70.5
	ONLINE-G	0.009	70.2	−0.067	65.3
	ONLINE-F	−0.204	64.0	−0.348	57.8
	JHU-SYNTAX	−0.261	62.4	−0.237	62.5
fi-en	ONLINE-B	**0.095**	66.9	**0.100**	65.4
	UEDIN-PBMT	**0.087**	66.3	**0.149**	66.6
	ONLINE-G	**0.084**	66.4	0.009	62.3
	UH-OPUS	**0.065**	65.9	**0.105**	65.3
	PROMT-SMT	−0.037	62.9	−0.093	58.8
	UEDIN-SYNTAX	−0.090	61.5	−0.041	60.9
	UH-FACTORED	−0.098	61.2	−0.020	61.1
	ONLINE-A	−0.126	60.6	−0.094	58.5
	JHU-PBMT	−0.391	52.7	−0.320	53.1
ro-en	ONLINE-B	**0.129**	73.9	0.051	66.7
	UEDIN-NMT	0.044	71.2	**0.258**	71.9
	UEDIN-PBMT	0.025	71.0	0.028	65.6
	UEDIN-SYNTAX	0.000	69.9	−0.020	64.6
	ONLINE-A	−0.012	69.7	−0.015	64.3
	LIMSI	−0.123	66.7	−0.071	62.8
	JHU-PBMT	−0.160	65.7	−0.187	60.2
ru-en	ONLINE-G	**0.115**	74.2	0.100	69.9
	AMU-UEDIN	**0.103**	73.3	0.178	72.2
	ONLINE-B	0.083	72.8	0.030	67.8
	NRC	0.060	72.7	0.092	69.9
	PROMT-RULE-BASED	0.044	72.1	−0.102	63.8
	UEDIN-NMT	0.011	71.1	**0.245**	74.3
	ONLINE-A	−0.007	70.8	0.020	66.7
	AFRL-MITLL-PHRASE	−0.040	70.1	0.047	68.4
	AFRL-MITLL-CONTRAST	−0.071	69.3	−0.020	66.5
	ONLINE-F	−0.322	61.8	−0.472	54.7
tr-en	ONLINE-B	**0.163**	57.1	**0.250**	60.0
	ONLINE-G	0.109	55.0	0.166	58.7
	ONLINE-A	0.002	52.2	0.130	57.8
	TBTK-SYSCOMB	−0.077	49.6	0.009	53.2
	PROMT-SMT	−0.079	49.2	−0.057	51.4
	YSDA	−0.088	49.5	−0.036	52.6
	JHU-PBMT	−0.355	41.0	−0.416	43.1
	JHU-SYNTAX	−0.364	40.8	−0.307	46.4
	PARFDA	−0.367	40.5	−0.406	42.3

DA to-English Translation Task

Table 9: DA mean scores for WMT16 translation task participating systems for translation into English.

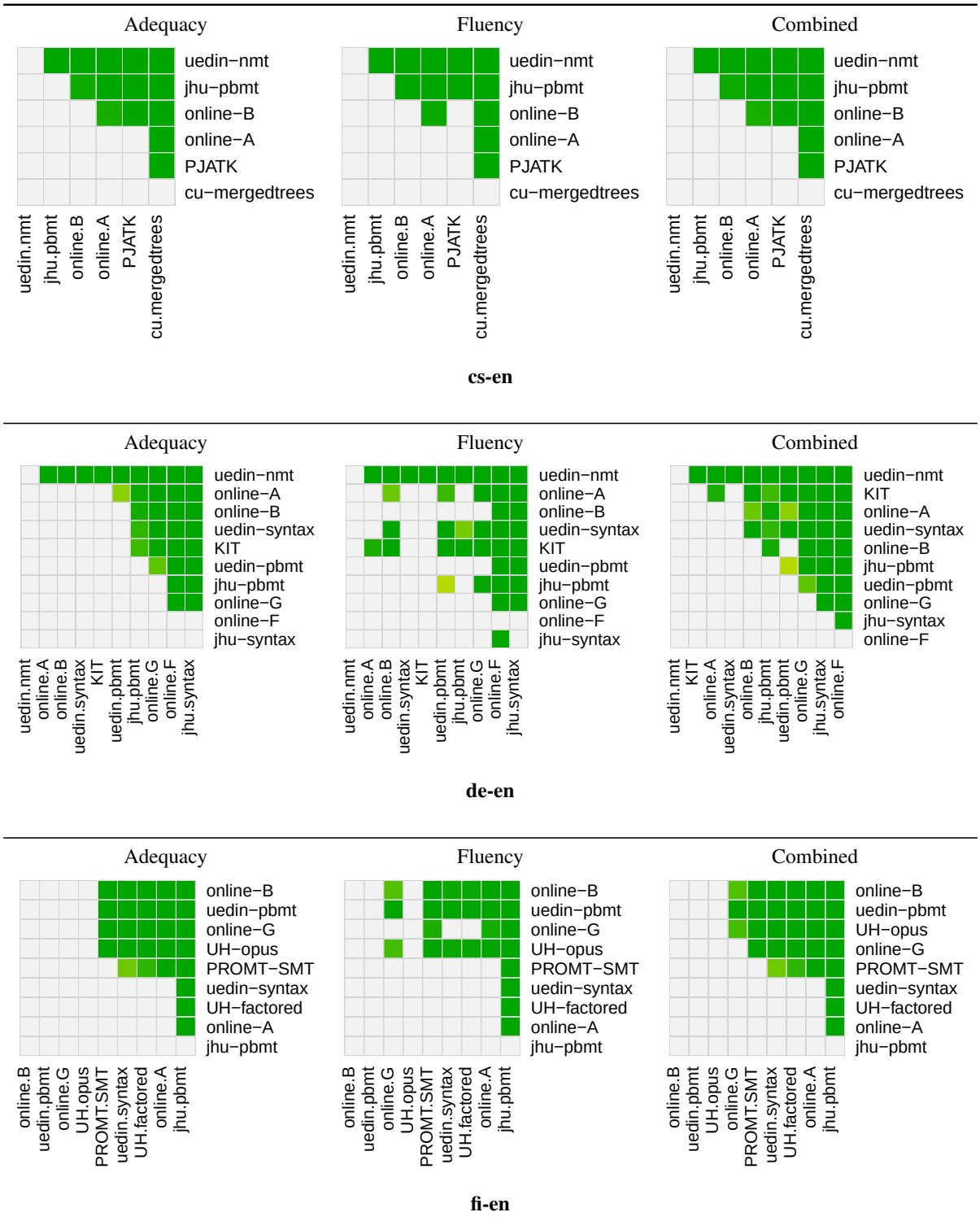

Figure 7: Significance test results for pairs of systems competing in the news domain translation task (cs-en, de-en, fi-en), where a green cell denotes a significantly higher DA adequacy or fluency score for the system in a given row over the system in a given column, "Combined" results show overall conclusions when adequacy is primarily used to rank systems with fluency used to break ties between systems tied with respect to adequacy.

Figure 8: Significance test results for pairs of systems competing in the news domain translation task (ro-en, ru-en, tr-en), where a green cell denotes a significantly higher DA adequacy or fluency score for the system in a given row over the system in a given column, "Combined" results show overall conclusions when adequacy is primarily used to rank systems with fluency used to break ties between systems tied with respect to adequacy.

	Adequacy	
	mean z	mean raw (%)
PROMT-RULE-BASED	**0.258**	69.0
ONLINE-G	0.101	63.8
ONLINE-B	0.092	62.5
AMU-UEDIN	0.084	63.4
UEDIN-NMT	0.062	63.2
ONLINE-A	−0.008	60.8
JHU-PBMT	−0.023	58.6
NYU-UMONTREAL	−0.042	58.3
LIMSI	−0.072	58.9
AFRL-MITLL-PHRASE	−0.077	58.3
AFRL-MITLL-VERB-ANN	−0.093	57.8
ONLINE-F	−0.489	43.7

DA English to Russian

Table 10: DA mean scores for WMT16 translation task participating systems for translation from English into Russian.

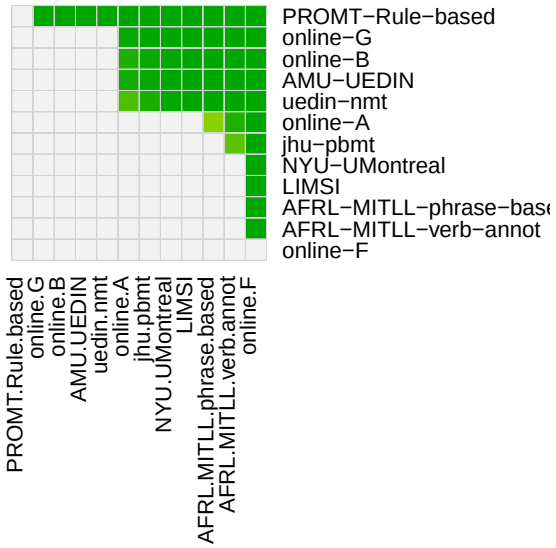

Figure 9: Significance test results for pairs of systems competing in the news domain translation task (en-ru), where a green cell denotes a significantly higher DA adequacy score for the system in a given row over the system in a given column.

cs-en	0.997
fi-en	0.996
tr-en	0.988
de-en	0.964
ru-en	0.961
ro-en	0.920
en-ru	0.975

DA Correlation with **RR**

Table 11: Correlation between overall DA standardized mean adequacy scores and RR Trueskill scores.

4 IT Translation Task

The IT-domain translation task introduced this year brought several novelties to WMT:

- 4 out of the 7 languages of the IT task are new in WMT (Bulgarian, Basque, Dutch and Portuguese),
- adaptation to the IT domain with its specifics such as frequent named entities (mostly menu items, names of products and companies) and technical jargon,
- adaptation to translation of answers in helpdesk service setting (many of the sentences are instructions with imperative verbs, which is very rare in the News translation task and may require adaptation of the whole translation pipeline, including e.g. part-of-speech taggers).

4.1 Data

The test set consisted of 1000 answers from the Batch 3 of the QTLeap Corpus.[10] The in-domain training data contained 2000 answers from the Batches 1 and 2 and also localization files from several open-source projects (LibreOffice, KDE, VLC) and bilingual dictionaries of IT-related terms extracted from Wikipedia. The out-of-domain training data contained all the corpora from the News Task (see Figure 1), plus PaCo2-EuEn Basque-English corpus and SETimes with Bulgarian-English parallel sentences.

"Constrained" systems were restricted to use only these training data provided by the organizers. Linguistic tools such as morphological analyzers, taggers, parsers, word-sense disambiguation or named entity recognizer were allowed in the constrained condition. The split of Batches 1 and 2 into the training set and development test set was left to the participants.

4.2 Submitted systems

31 systems were submitted in total for the 7 language pairs.

Avramidis (2016) describes all English→German QTL-* systems (DFKI). Rosa et al. (2016) describe QTL-CHIMERA (Charles University). Gaudio et al. (2016) describe the remaining QTL-* systems (partners from the QTLeap project: HF&FCUL for Portuguese, UPV/EHU for Spanish and Basque, IICT-BAS for Bulgarian, CUNI for Czech and UG for Dutch). Duma and Menzel (2016) describe UHDS-DOC2VEC and UHBS-LMI (University of Hamburg). Pahari et al. (2016) describe JU-USAAR (Jadavpur University & Saarland University). Cuong et al. (2016) describe ILLC-UVA-SCORPIO (University of Amsterdam). IILC-UVA-DS is based on Hoang and Sima'an (2014). PROMT-RULE-BASED and PROMT-HYBRID systems were submitted by the PROMT LLC company and they are not described in any paper.

QTL-MOSES is the standard Moses setup (MERT-tuned on the in-domain training data, but otherwise without any domain-adaptation) and serves as a baseline.

4.3 Human evaluation

The main results are presented in Table 12. The PROMT-* systems won all three language pairs, for which they were submitted, but they were trained using additional training data not available to other participants, so they are considered unconstrained and not comparable to the constrained systems. In all language pairs except for English→Bulgarian, the baseline (QTL-MOSES) was outperformed by all other systems.

Table 13 reports the amount of pairwise comparisons collected and inter- and intra-annotator agreement of the human evaluation, which is in a similar range as in the News task (cf. Tables 4 and 5).

5 Biomedical Translation Task

This is the first time that we have run the Biomedical Translation task at WMT. This task aims to evaluate systems for the translation of biomedical titles and abstracts from scientific publications. In this first edition of the challenge, we have focused on three language pairs (considering both translation directions), namely, English/Portuguese (EN/PT), English/Spanish (EN/ES) and English/French (EN/FR), and documents in the two sub-domains of biological sciences and health sciences.

5.1 Task description

The participants were provided with training data and were required to submit automatic translations

[10] http://metashare.metanet4u.eu/go2/qtleapcorpus

English→Bulgarian

#	score	range	system
1	5.26	1	QTL-MOSES
2	−5.26	2	QTL-DEEPFMOSES

English→Czech

#	score	range	system
1	0.53	1–2	QTL-CHIMERA-PURE
	0.43	1–2	ILLC-UVA-DS
2	0.13	3	QTL-TECTOMT
3	−0.47	4–5	QTL-CHIMERA-PLUS
	−0.62	4–5	QTL-MOSES

English→German

#	score	range	system
1	1.61	1	PROMT-RULE-BASED
2	−0.04	2–5	UHBS-LMI
	−0.06	2–6	UHDS-DOC2VEC
	−0.06	2–6	QTL-RBMT-SMTMENUS
	−0.09	3–6	RBMT
	−0.10	3–6	QTL-RBMT-MENUS
3	−0.19	7–8	DFKI-SYNTAX
	−0.19	7–8	JU-USAAR
4	−0.38	9	QTL-SELECTION
5	−0.49	10	QTL-MOSES

English→Spanish

#	score	range	system
1	3.53	1	PROMT-HYBRID
2	−0.80	2–3	QTL-CHIMERA
	−0.81	2–3	QTL-TECTOMT
3	−1.93	4	QTL-MOSES

English→Basque

#	score	range	system
1	1.57	1	QTL-TECTOMT
2	−1.57	2	QTL-MOSES

English→Dutch

#	score	range	system
1	1.95	1	ILLC-UVA-SCORPIO
2	0.36	2	QTL-CHIMERA
3	0.15	3	QTL-TECTOMT
4	−2.46	4	QTL-MOSES

English→Portuguese

#	score	range	system
1	4.61	1	PROMT-HYBRID
2	−1.06	2	QTL-TECTOMT
3	−1.27	3	QTL-CHIMERA
4	−2.28	4	QTL-MOSES

Table 12: Official results for the WMT16 IT translation task. Systems are ordered by their inferred system means, though systems within a cluster are considered tied. Lines between systems indicate clusters according to bootstrap resampling at p-level $p \leq .05$. Systems with gray background indicate use of resources that fall outside the constraints provided for the shared task.

Language pair	Systems	Comparisons	Comparisons/sys	Inter-κ	Intra-κ
English→Bulgarian	2	1,769	884.5	0.447	0.627
English→Czech	5	16,870	3,374.0	0.330	0.463
English→German	10	38,733	3,873.3	0.385	0.492
English→Spanish	4	8,538	2,134.5	0.351	0.398
English→Basque	2	1,485	742.5	0.483	0.610
English→Dutch	4	7,278	1,819.5	0.258	0.249
English→Portuguese	4	7,794	1,948.5	0.594	0.705
Sum	31	82,467			
Mean			2,660.2	0.407	0.506

Table 13: Amount of manual-evaluation pairwise comparisons (after "de-collapsing" *multi-system outputs*) collected and κ scores measuring inter- and intra-annotator agreement in the IT task. Cf. Tables 3, 4 and 5 for the respective News task statistics.

for each document in the test set. Details on the data, baseline system, automatic evaluation and manual validation are described below.

Data

We provided the participants with training data of parallel documents for the three language pairs as well as monolingual documents for each of the four languages, as summarized in Table 14. We did not provide any development data and the participants were free to split the training data into a training and a development datasets.

The training data consisted mainly of the Scielo corpus (Neves et al., 2016), a parallel collection of scientific publications composed of either titles, abstracts or title and abstracts which were retrieved from the Scielo database. For the Scielo corpus, we compiled parallel documents for all language pairs in the two sub-domains, except for the EN/FR, where only health was considered, as there were inadequate parallel documents available for biology in that pair. In previous work (Neves et al., 2016), the training data was aligned using the GMA alignment tool. The quality of the alignment was found to be satisfactory so that aligned training data could be made available to the participants.

The test set consisted of 500 documents (title and abstract) for each of the two directions of each language pair, i.e., English to Portuguese (en-pt), Portuguese to English (pt-en), English to Spanish (en-es), Spanish to English (es-en), English to French (en-fr) and French to English (fr-en). None of the test documents was included in the training data and there is no overlap of documents between the test sets for any language pair, translation direction and sub-domain.

Additionally, we prepared a corpus of parallel titles from MEDLINE® for all three language pairs. Finally, we also provided monolingual documents for the four languages, i.e., English, French, Spanish and Portuguese, retrieved from the Scielo database. These consist of documents in the Scielo database which have no corresponding document in another language.

Evaluation metric

We computed the BLEU score for each of the runs in comparison to the reference translation, i.e., the original text made available in the Scielo database, as provided by the authors of the publications.

Baseline

Our baseline system was described in previous work (Neves et al., 2016). It consists of the statistical MT system Moses [11] trained on both the Scielo corpus and on the parallel collection of Medline titles. We did not make use of the monolingual collection as we did not train a language model.

Manual validation

We carried out a manual evaluation for 100 random sentences for some selected pairs in the test data. We used the 3-way ranking task in the Appraise tool [12] which typically shows the source and the reference translation, and allows the pairwise comparison of two translations (A and B).

However, to distance the manual evaluation from the automatic BLEU evaluation which compares automatic runs to the reference translation, we treated the reference translation as one of the systems and therefore suppressed the reference translation in the interface. Evaluators were only presented with the source sentence, and two translations to rank. Evaluators were blind to the nature of the sentences they were evaluating: automatic system A vs. system B, reference translation vs. system, or system vs. reference translation.

When comparing two translations in the 3-way ranking task in Appraise, evaluators were presented with four options: (1) A>B, translation A is better than translation B; (2) A=B, the quality of the two candidate translations is similar; (3) A<B, translation B is better than translation A; and (4) Flag Error, to indicate that one of the translations did not seem to refer to the same source sentence or there is some other misalignment. The latter situation could happen when the original sentence pairs were not perfectly aligned. This may be due to the fact that the reference translations are created by the article authors independently of the WMT challenge goals. These authors are not professional writers or professional translators, so that some of the content may only be present in one of the languages, i.e., not every sentence in one language has a directly corresponding sentence in the other language. Thus, when selecting the corresponding sentences in the reference translation, we do it based on the automatic alignment provided by the GMA tool, which performs with at least 80% accuracy for our training data (Neves

[11] http://www.statmt.org/moses/

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

Table 14: Statistics on training and test collections for the Biomedical Translation Task. "T" corresponds to percentage of titles and "A" to percentage of abstracts, separated by a slash. "Docs" to total number of documents, "Lang" identifies the language,"Sents" to total number of sentences and "Tokens" to total number of tokens.

Dataset	Train	Docs	T/A	Lang	Sents	Tokens
Biological	EN/ES	17,672	49.4/97.7	EN	138,073	3,819,190
				ES	128,894	3,887,818
	EN/PT	18,180	31.1/96.1	EN	128,357	3,807,296
				PT	125,717	3,598,618
Health	EN/ES	75,856	55.6/99.5	EN	628,966	15,978,198
				ES	606,231	17,168,994
	EN/PT	65,659	74.0/92.8	EN	541,272	14,457,939
				PT	525,721	14,447,017
	EN/FR	1,135	64.5/99.7	EN	9,393	250,907
				FR	9,501	320,132
Dataset	Test	Docs	T/A	Lang	Sents	Tokens
Biological	en-es	500	100/100	EN	4,344	116,388
				ES	4,070	125,491
	es-en	500	100/100	ES	4,113	124,343
				EN	4,405	115,045
	en-pt	500	100/100	EN	4,333	114,705
				PT	4,205	120,591
	pt-en	500	100/100	PT	4,029	114,970
				EN	4,164	108,120
Health	en-fr	500	100/100	EN	5,093	137,321
				FR	5,782	208,795
	fr-en	500	100/100	FR	5,784	206,559
				EN	5,178	137,638
	en-es	500	100/100	EN	5,111	127,112
				ES	5,027	141,473
	es-en	500	100/100	ES	5,198	144,666
				EN	5,276	128,742
	en-pt	500	100/100	EN	3,858	99,001
				PT	3,776	101,991
	pt-en	500	100/100	PT	3,826	106,735
				EN	3,930	102,813

et al., 2016).

Regarding assigning the second option, i.e., A=B, we considered situations in which both translations were equally bad or good. In some cases, both candidate translations exhibited either lexical or grammatical issues, but the evaluator could not rank one candidate as definitely better or worse than the other. Sometimes, both candidates were correct and were acceptable translations of the source sentence, even if not identical. Currently, this distinction is not captured in the statistics computed by Appraise.

5.2 Participants

Five teams participated in the Biomedical Translation task, submitting a total of 40 runs. Participants are listed in Table 15; a short description of their systems is provided below.

Istrionbox The Istrionbox team utilized a non-log-linear model based on a weighted average of the translation and language models. They aligned the training documents on the phrase level using an aligner based on a lexicon which contains more than 930,000 terms derived from many parallel corpora for English/Portuguese. The language model was based on phrases, instead of words, as well as the translation model. For the various runs that the team submitted, they experimented with assigning equal or different weights for the distinct models trained on the biological or the health corpora, and they also considered a bilingual lexicon and named entities.

IXA The IXA team adapted a general-domain statistical machine translation system to the biomedical domain. Three approaches were developed for English-Spanish and Spanish-English language pairs, using Moses and three corpora (News corpora, Scielo Health and Scielo Biological, both the bilingual and monolingual documents). In the system used for the first submission, the medical vocabulary SNOMED-CT is used to extend the vocabulary to address the problem of out-of-vocabulary (OOV) words. In the system used for the second submission, OOV words are

Team ID	Participating team
Istrionbox	Istrionbox, Portugal (Aires et al., 2016)
IXA	University of the Basque Country UPV/EHU, Spain (Perez-de Viñaspre and Labaka, 2016)
LIMSI-TLP	LIMSI, France (Ive et al., 2016)
TALP-UPC	Universitat Politècnica de Catalunya, Spain (Costa-jussà et al., 2016)
uedin	University of Edinburgh, UK (Williams et al., 2016)

Table 15: Participants in the WMT16 Biomedical Translation task.

addressed by expanding generated phase tables with morphological variants and transliterations of the remaining words. In the system used for the third submission, the IXA team used the test set provided by the organizers to optimize the method used in the second submission.

TALP The TALP team's system is a standard phrase-based system based on Moses and MERT and enhanced with vocabulary expansion using bilingual word embeddings and a character-based neural language model with rescoring. The former focuses on resolving out-of-vocabulary words, while the latter enhances the fluency of the system.

LIMSI-TLP The LIMSI-TLP system is a MOSES-based statistical machine translation system, rescored with Structured Output Layer neural network models. It relied on additional in-domain data, including data from the WMT'14 medical translation task (English-French) and a set of English-French Cochrane systematic review abstracts. They also experiment with a confusion network system combination which combines the outputs of Phrase Based SMT systems trained either to translate entire source sentences or specific syntactic constructs extracted from those sentences. The approach is implemented using Confusion Network decoding.

uedin The University of Edinburgh team used the phrase-based statistical model from Moses including hierarchical lexicalized reordering model with four orientations in both directions. The translation model was trained on data from the WMT13, the Scielo training data as well as the EMEA corpus. The language model was based on the interpolation of various language models trained separately on monoligual English corpora, such as the WMT14 medical, Scielo, EMEA and English LDC GigaWord corpus.

5.3 Results

The five participating teams submitted a total of 40 runs. However, only the Spanish–English and English–Spanish language pairs attracted submissions from more than one team. In addition, one language pair (fr-en) did not receive any submission. Table 16 presents the BLEU score for each run as well as for our baseline system.

All runs obtained a much higher BLEU score than the baseline system, except for the en-pt and pt-en submissions, with BLEU scores just slightly superior to the baseline. The LIMSI run showed the best improvement over the baseline (246% absolute improvement, from 9.24 to 22.75). Overall, however, the BLEU scores for all language pairs remain quite moderate. Regarding comparison of the various runs and teams for each language pair, we did not observe considerable differences between them, except for the the runs of the "uedin" system, which obtained around two BLEU points more than other runs.

We rank the systems as follows according to their BLEU scores, with B=biology and H=health, and bl=baseline:

- en-pt(B): Istrionbox>bl;
- en-pt(H): Istrionbox>bl;
- pt-en(B): Istrionbox>bl;
- pt-en(H): Istrionbox>bl;
- en-es(B): TALP>IXA>bl;
- en-es(H): TALP>IXA>bl;
- es-en(B): uedin>IXA>TALP>bl;
- es-en(H): uedin>IXA>TALP>bl;
- en-fr(H): LIMSI>bl;

Languages	Team ID	Run ID	BLEU score	
			Biological	Health
en-pt	Istrionbox	1	17.55	19.01
		2	16.47	18.33
		3	16.45	18.37
	Baseline	-	15.38	17.22
pt-en	Istrionbox	1	20.88	21.50
		2	20.17	20.17
		3	20.14	20.62
	Baseline	-	17.59	18.48
en-es	IXA	1	31.57	28.09
		2	31.32	28.06
		3	29.61	28.13
	TALP	1	31.18	28.11
		2	31.17	27.85
		3	33.22	29.47
	Baseline	-	17.82	16.88
es-en	IXA	1	30.66	27.96
		2	30.59	27.97
		3	29.51	28.12
	TALP	1	29.68	27.42
		2	29.41	26.74
		3	29.83	27.27
	uedin	1	31.49	29.05
	Baseline	-	18.78	16.92
en-fr	LIMSI	1	-	22.52
		2	-	22.75
	Baseline	-	-	9.24

Table 16: Official BLEU scores for the WMT16 Biomedical Translation task.

For the pairwise manual validation of sentences, and given the high number of runs for some language pairs, e.g., Spanish–English and English–Spanish, we did not perform a pairwise evaluation for every pair of two systems. Instead, we considered only one run from each participant for each language pair and dataset: the one that achieved the best BLEU score in the automatic evaluation. An exception was made for the English–French and English-Portuguese tasks for which we had only one participating team: we considered all combinations of runs and reference translations for English–French and combinations of the reference translation and both the run with best BLEU score and the one that the participant (Istrionbox) reported as their best run. The results of the manual validation are presented in Table 17.

Only one run (IXA run 3, English–Spanish, health dataset) was comparable to the reference translation: 30 vs. 26 for A>B and A<B, respectively. For all other cases, the reference translation was assigned to be better than the other translation at least twice as many times.

Regarding comparison between teams and runs, i.e., ES2PT (biological and health) and English–French, we did not observe much difference when comparing distinct runs of the same team. When comparing runs from distinct teams, IXA clearly outperformed TALP in two comparisons: Spanish–English biological (57 vs. 24) and Spanish–English health (48 vs. 22). On the other hand, TALP slightly outperformed IXA in one dataset: English–Spanish biological (16 vs. 7). Finally, the uedin system was clearly superior to TALP in the Spanish–English biological dataset (60 vs. 20) and to both TALP and IXA in the Spanish–English health dataset (54 vs. 19 and 41 vs. 15, respectively).

We rank the systems as follows according to our manual validation (ref=reference):

Datasets	Pairs	Runs	Total	A>B	A=B	A<B
Biological	en-es	TALP run3 vs. reference	97	18	20	59
		IXA run1 vs. TALP run3	70	7	47	16
		reference vs. IXA run1	96	50	30	16
	es-en	IXA run1 vs. reference	76	17	19	40
		reference vs. uedin run1	75	43	14	18
		TALP run3 vs. IXA run1	100	24	19	57
		reference vs. TALP run3	68	52	6	10
		IXA run1 vs. uedin run1	100	30	31	39
		uedin run1 vs. TALP run3	100	60	20	20
	en-es	reference vs. Istrionbox run1	80	54	20	6
		Istrionbox run3 vs. Istrionbox run1	99	22	52	25
		Istrionbox run3 vs. reference	80	4	14	62
	pt-en	reference vs. Istrionbox run3	78	67	7	4
Health	en-fr	reference vs. LIMSI-TLP run2	91	71	5	15
		LIMSI-TLP run1 vs. LIMSI-TLP run2	88	26	40	22
		LIMSI-TLP run1 vs. reference	85	8	12	65
	en-es	reference vs. IXA run3	93	30	37	26
		IXA run3 vs. TALP run3	82	23	40	19
		TALP run3 vs. reference	94	21	28	45
	es-en	reference vs. IXA run3	82	41	29	12
		IXA run3 vs. TALP run1	100	48	30	22
		TALP run1 vs. reference	75	8	20	47
		IXA run3 vs. uedin run1	100	15	44	41
		reference vs. uedin run1	79	44	20	15
		TALP run1 vs. uedin run1	100	19	27	54
	en-pt	Istrionbox run3 vs. Istrionbox run1	100	29	42	29
		Istrionbox run1 vs. reference	80	4	15	61
		reference vs. Istrionbox run3	82	62	17	3
	pt-en	Istrionbox run1 vs. reference	89	6	1	82

Table 17: Results for the manual validation carried out in Appraise for the Biomedical Translation task.

- en-pt (B): ref>Istrionbox;
- en-pt (H): ref>Istrionbox;
- pt-en (B): ref>Istrionbox;
- pt-en (H): ref>Istrionbox;
- en-es (B): ref>TALP> IXA;
- en-es (H): {IXA,ref}>TALP;
- es-en (B): ref>uedin>IXA>TALP;
- es-en (H): ref>uedin> IXA>TALP;
- en-fr (H): ref>LIMSI;

5.4 Discussion

In this section we analyze the errors we observed in the translations submitted by teams, the lessons we learned in this first edition of the task and our plans for future work.

Error analysis. During our manual analysis of a sample of the translations that were submitted for the test data, we noticed that their quality is still poor in comparison to the reference translations. We identified numerous problems, as summarized below:

- many missing words or words in the source language mixed in with the target language, probably due to words or concepts in the source language that could not be translated to the target language;
- incorrect ordering of adjectives and nouns, given that, in contrast to English, nouns typically precede adjectives in Portuguese, Spanish and French;
- incorrect agreement of nouns, verbs and adjectives with respect to gender and number;
- incorrect punctuation, e.g., periods placed in the middle of a sentence;
- incorrect casing for words, e.g., common words which were capitalized or in upper case;
- missing translations for acronyms, i.e., the acronym in the source language was used instead.

We note that some of these issues were ignored during the manual evaluation, for instance, incorrect capitalization was not penalized if the translation was otherwise better or comparable to the other translation.

Lessons learned. We performed a comparison of the systems based only on the overall results on the complete test set and on the samples of sets that we randomly selected for manual validation. For this first edition of the Biomedical Translation task, we aimed at providing an evaluation platform for the automatic translation of scientific publications, in particular for titles and abstracts in the biomedical domain.

In this first edition of the task, the training and test data was obtained from the parallel publications available in Scielo. We did not perform manual translation of the documents for either the training or the test data, but rather used the original text available in Scielo for all languages under consideration here. In practice, this means that the reference translations were produced by the article authors independently of the WMT challenge goals. These authors are not professional writers or professional translators, and some of them may have limited proficiency in the languages they are required to use for publication. This situation has an impact on the quality of the reference translations, compared to other WMT tasks. It is reflected in the manual evaluation which indicates that for some language pairs (notably English–Spanish health), participant runs were rated overall as better or equal to the reference translation. Our experience with this first edition of the task indicates that the Scielo corpus is a valuable resource for biomedical WMT, however more work is needed in terms of quality assurance to ensure that meaningful evaluation results can be obtained.

Plan for future editions. In next editions, we plan to build on the established pipeline to collect and pre-process Scielo data to prepare a new test dataset. More importantly, we plan to work towards improved data and evaluation quality.

While we initially focused on characterizing the quality of the alignment in the parallel Scielo corpus, we are planning to craft a higher quality dataset by removing any sentence pairs with alignment issues. Furthermore, the data set will also be pruned for sentences exhibiting lexical, grammatical or fluency issues. These steps will contribute to improve the significance of the evaluation results, especially in terms of BLEU scores.

Furthermore, we believe that the nature of scientific texts and biomedical texts in particular calls for specific evaluation metrics. One of the intended uses of translation systems in the biomedical domain is to provide health professionals with access to the latest research results that are published in a language other than their native language. Consequently, health professionals may use the translated information to make clinical decisions impacting patients care. It is vital that translation systems do not contribute to the dissemination of incorrect clinical information. Therefore, the evaluation of biomedical translation systems should include an assessment at the document level indicating whether a translation conveyed erroneous clinical information.

6 Quality Estimation

The fifth edition of the WMT shared task on quality estimation (QE) of machine translation (MT) builds on the previous editions of the task (Callison-Burch et al., 2012; Bojar et al., 2013, 2014, 2015), with "traditional" tasks at sentence and word levels, a new task for entire documents quality prediction, and a variant of the word-level task: phrase-level estimation.

The goals of this year's shared task were:

- To advance work on sentence and word-level quality estimation by providing domain-specific, larger and professionally annotated datasets.
- To analyse the effectiveness of different types of quality labels provided by humans for longer texts in document-level prediction.
- To investigate quality estimation at a new level of granularity: phrases.

These goals are addressed through three groups of tasks: Task 1 at sentence level (Section 6.3), Task 2 at word and phrase levels (Section 6.4), and Task 3 at document level (Section 6.6). Tasks 1 and 2 provide the same dataset with English-German translations generated by a statistical machine translation (SMT) system, while Task 3 provides an English-Spanish dataset of translations taken from all participating systems in WMT08-WMT13. These datasets were annotated with different labels for quality: for Tasks 1 and 2, the labels were automatically derived from the post-editing of the machine translation output, while for Task 3, scores were computed based on a two-stage post-editing process. Any external resource, including additional quality estimation training data, could be used by participants

(no distinction between *constrained* and *unconstrained* tracks was made). As presented in Section 6.1, 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, with suggested methods for all levels of prediction. Participants, described in Section 6.2, could submit up to two systems for each task.

Data used to build MT systems or internal system information (such as model scores or n-best lists) were made available on request for Tasks 1 and 2.

6.1 Baseline systems

Sentence-level baseline system: For Task 1, `QuEst++`[13] (Specia et al., 2015) was used to extract 17 features from the SMT source/target language training corpus:

- Number of tokens in source & 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 probability of source and target sentences based on models built from the SMT training corpus.
- Average number of translations per source word in the sentence as given by IBM Model 1 extracted from the SMT training corpus.
- 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 SMT training corpus.
- Percentage of unigrams in the source sentence seen in the source SMT training corpus.

These features were used to train a Support Vector Regression (SVR) algorithm using a Radial Basis Function (RBF) kernel within the `scikit-learn` toolkit (Pedregosa et al., 2011).[14]

[13] `https://github.com/ghpaetzold/questplusplus`

[14] `http://scikit-learn.org/`

The γ, ϵ and C parameters were optimised via grid search with 5-fold cross validation on the training set.

Word-level baseline system: For Tasks 2 and $2p$, the baseline features were extracted with the `Marmot` tool (Logacheva et al., 2016b).

For the baseline system we used a number of features that have been found the most informative in previous research on word-level QE. Our baseline set of features is loosely based on the one described in (Luong et al., 2014). It contains the following 22 features:

- Word count in the source and target sentences, 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 one word.
- Source word aligned to the target token, its left and right contexts of one word. The alignments were taken from the SMT system that produced the automatic translations.
- Binary dictionary features: whether target token is a stopword, a punctuation mark, a proper noun, a number.
- Target language model features:
  - The order of the highest order ngram which starts and end with the target token.
  - 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 (the backoff behaviour was computed as described in (Raybaud et al., 2011)).
- The order of the highest order ngram which starts and ends with the source token.
- The Part-of-speech tags of the target and source tokens.

This set of baseline features is similar to the one used at WMT15 QE shared task (Bojar et al., 2015). We excluded three features used the last

year: pseudo-reference features and number of WordNet senses for the source and target tokens.

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

Phrase-level baseline system: The phrase-level features were also extracted with `Marmot`, but they are different from the word-level features. The baseline set of phrase-level features is based on a list of features which were used for sentence-level QE in `QuEst++` toolkit. These so-called "black-box" features do not use the internal information from the MT system. We use the following feature set consisting of 72 features, using the SMT source/target language training corpus:

- Source phrase frequency features:
  - average frequency of ngrams (unigrams, bigrams, trigrams) from different quartiles of frequency (from the low frequency to high frequency ngrams);
  - percentage of distinct source ngrams (unigrams, bigrams, trigrams) seen in a corpus of the source language.
- Translation probability features:
  - average number of translations per source word in the sentence (with different translation probability thresholds: 0.01, 0.05, 0.1, 0.2, 0.5);
  - average number of translations per source word in the sentence (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 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 and the source.
- Language model features:
  - log probability of the source and the target phrases;
  - perplexity of the source and the target phrases.
- Phrase statistics:
  - lengths of the source and target phrases;
  - ratio of the source and the target phrase lengths;
  - average length of tokens in source and target phrases;
  - average occurrence of target word within the phrase.
- Alignment features:
  - Number of unaligned target words;
  - Number of target words aligned to more than one word;
  - Average number of alignments per word in the target phrase.
- Part-of-speech features:
  - percentage of content words in the source and target phrases;
  - percentage of words of a particular part-of-speech (verb, noun, pronoun) in the source and the target phrases;
  - ratio of numbers of words of a particular part-of-speech (verb, noun, pronoun) in the source and the target phrases;
  - percentage of numbers and alphanumeric tokens in the source and the target phrases;
  - ratio of the percentage of numbers and alphanumeric tokens in the source and the target phrases;

This feature set was originally designed for sentences. We expect that since phrases are sequences of words of varied length, they can be treated analogously for QE. However, unlike sentences, which are translated independently, phrases are related to their neighbouring phrases in a sentence, and in this respect they are similar to words in the context of QE. Therefore, as in 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` using the passive-aggressive optimisation algorithm.

Document-level baseline system: For Task 3, 17 baseline features equivalent to those for sentence level were extracted at document level using `QuEst++`. These features are aggregations of sentence-level baseline features. Some sentence-level features were summed (number of tokens in the source and target sentences and number of punctuation marks in source and target sentences), while all remaining were averaged.

The model was trained with a SVR algorithm with RBF kernel using the `scikit-learn` toolkit. The γ, ϵ and C parameters were optimised via grid search with 5-fold cross validation on the training set.

6.2 Participants

Table 18 lists all participating teams submitting systems to any of the tasks. Each team was allowed up to two submissions for each task. In the descriptions below, participation in specific tasks is denoted by a task identifier.

CDACM (Task 2): The CDACM team participated in Task 2 for the word and phrase-level QE. They use a Recurrent Neural Network Language Model (RNN-LM) architecture for word-level QE. To estimate the phrase-level quality, they use the output of the word-level QE system. For this task, they use a modified RNN-LM with other RNN variants like Long Short Term Memory (LSTM), deep LSTM and Gated Recurring Units (GRU). The modified system predicts a label (OK/BAD) rather than predicting the word as in the case of standard RNN-LM. The input to the system is a word sequence, similar to the standard RNN-LM. They also tried bilingual models with RNN-LM and found that they perform better than monolingual models. In the training data, the distribution of labels (OK/BAD) is skewed, with significantly more OK labels. To handle this issue, they use strategies to replace the OK label with sub-labels to balance the distribution. The sub-labels are OK_B, OK_I, OK_E, depending on the location of the token in the sentence.

POSTECH (Task 1, Task 2): POSTECH's submissions (SENT/RNN for Task 1, WORD/RNN for Task 2 and PHR/RNN for Task 2p) are RNN-based QE systems consisting of two component: two bidirectional RNNs on the source and target sentences in the first component and other RNNs for predicting the final quality in the second component. The first component is an RNN-based modified neural MT model which generates quality vectors. Quality vectors indicate a sequence of vectors about target words' translation quality. The second component using other RNNs predicts the quality at sentence level (Task 1), word level (Task 2), and phrase level (Task 2p). POSTECH's RNN-based systems are entirely neural approaches for QE. Due to the small amount of data to train the prediction models, each component of the systems is trained separately by using different training data. To train the first component of the systems, the Europarl v7 English-German parallel corpus was used. To train the second component of the systems, WMT16 QE task English-German datasets were used.

RTM (Task 1, Task 2, Task 3): Referential translation machines (RTMs) (Biçici and Way, 2015) are a language-independent approach for predicting translation quality, as well as for addressing other text similarity tasks. They eliminate the need to access any task or domain specific information or resource. SVR and regression trees are used in combination with feature selection and partial least squares for the document and sentence-level prediction tasks and global linear models with dynamic learning were used for the word and phrase-level prediction tasks.

SHEF (Task 1): The SHEF systems exploit RNNs and the principle of compositionality to offer a resource-light solution to sentence-level QE. They use only one side of the translation, the source (SRC) or the target (TGT). They split the sentence in ngrams and train a model that predicts the quality of ngrams. To calculate the quality of an entire sentence translation, they split its source/target side in ngrams, estimate their quality individually, then average their quality scores. They use word embedding models trained over 7 billion words as external resource (English and German) using `word2vec`.

SHEF-LIUM (Task 1): The two joint submissions from the University of Sheffield and LIUM use (i) a Continuous Space Language Model (CSLM) to extract sentence embeddings and cross-entropy scores, (ii) a neural network MT (NMT) model, (iii) a set of `QuEst++` features (iv) a combination of features produced by `QuEst++` and the features produced with CSLM and NMT. When added to `QuEst++` standard feature sets for Task 1, the CSLM sentence embed-

ID	Participating team
CDACM	Centre for Development of Advanced Computing, India (Patel and M, 2016)
POSTECH	Pohang University of Science and Technology, Republic of Korea (Kim and Lee, 2016)
RTM	Referential Translation Machines, Turkey (Bicici, 2016b)
SHEF	University of Sheffield, UK (Paetzold and Specia, 2016)
SHEF-LIUM	University of Sheffield, UK and Laboratoire d'Informatique de l'Université du Maine, France (Shah et al., 2016)
SHEF-MIME	University of Sheffield, UK (Beck et al., 2016)
UAlacant	University of Alicante, Spain (Esplà-Gomis et al., 2016)
UFAL	Nile University, Egypt & Charles University, Czech Republic (Abdelsalam et al., 2016)
UGENT	Ghent University, Belgium (Tezcan et al., 2016)
UNBABEL	Unbabel, Portugal (Martins et al., 2016)
USFD	University of Sheffield, UK (Logacheva et al., 2016a)
USHEF	University of Sheffield, UK (Scarton et al., 2016)
UU	Uppsala University, Sweden (Sagemo and Stymne, 2016)
YSDA	Yandex School of Data Analysis, Russia (Kozlova et al., 2016)

Table 18: Participants in the WMT16 Quality Estimation shared task.

ding features along with the cross entropy and NMT likelihood led to large improvements in prediction, and achieved third place in the scoring and second place in the ranking task variants according to the official evaluation metrics. Neural network features alone also performed very well. This is a very encouraging finding since for many language pairs it is sometime hard to find appropriate resources to build hand-crafted features, while the neural network features used only require (sufficient) monolingual data to train models, which is available in abundance for many languages.

SHEF-MIME (Task 2): The University of Sheffield's submission to the word-level QE task is based on imitation learning, an approach that treats structured prediction as a sequence of actions taken by a binary classifier. This approach allows the use of arbitrary information from previous tag predictions and has the ability to train the classifier using non-decomposable loss functions over the predicted structure. The submitted system uses the baseline features provided by the shared task organisers plus additional features relying on the predicted structure, such as previous tag ngrams and the total number of BAD predictions. It employs an online learning algorithm as the underlying classifier and uses a loss function based on the official shared task evaluation metric. No external data or resources were used for this submission.

UALacant (Task 2): The submissions of the Universitat d'Alacant team focus for Task 2 were obtained by applying the approach by Esplà-Gomis et al. (2015), which uses any source of bilingual information available online in order to spot sub-segment correspondences between the source segment and the translation hypothesis. These sub-segment correspondences are used to extract a collection of features that are then used by a multilayer perceptron to determine the final word-level QE labels. The probabilities provided by this classifier for every word in a phrase are then used as new features for a second multilayer perceptron that is able to obtain quality estimates at the phrase level. Three sources of bilingual information available online were used by the UAlacant submissions: two online MT systems, `Lucy LT KWIK`[15] and `Google Translate`,[16] and the bilingual concordancer `Reverso Context`.[17] Two systems were submitted, both for word-level and phrase-level QE tasks: one using only features based on external sources of bilingual information, and another combining them with the baseline features provided by the task organisers.

[15] `http://www.lucysoftware.com/catala/traduccio-automatica/kwik-translator-/`

[16] `http://translate.google.com`

[17] `http://context.reverso.net/translation/`

UFAL (Task 1): The submission is based on word alignments and bilingual distributed representations to introduce a new set of features for the sentence-Level QE task. The features extracted include three alignment-based features, three bilingual embedding-based features, two embedding-based features constrained on alignment links, as well as a set of 74 bigrams used as boolean features. The set of bigrams represents the most frequent bigrams in translations that have changed after the post-edition, and they are compiled by aligning translations to their post-editions provided in the WMT QE datasets. To produce these features, `GIZA++` (Och and Ney, 2003) was used for word alignment and `Multivec` (Berard et al., 2016) was used for the bilingual model, which jointly learns distributed representations for source and target languages using a parallel corpus. To build the bilingual model, domain-specific data compiled from the resources made available for the WMT 16 IT-Domain shared task was used. As prediction model, a Linear Regression model using `scikit-learn` was built using a combination of `QuEst++` baseline features and the new features proposed.

UGENT-LT3 (Task 1, Task 2): The submissions for the word-level task use 41 features in combination with the baseline feature set to train binary classifiers. The 41 additional features attempt to capture accuracy errors (concerned with the meaning transfer from the source to target sentences) using word and phrase alignment probabilities, fluency errors (concerned with the well-formedness of target sentence) using language models trained on word surface forms and on part-of-speech tags, and terminology errors (concerned with the domain-specific terminology) using a bilingual terminology list. Based on the combined feature set, SCATE-RF uses random forests for binary classification, which combines decision trees into an ensemble. SCATE-ENS uses the same feature set and combines different algorithms into an ensemble by applying the majority voting scheme. For the sentence-level task, SCATE-SVM1 adds 18 features to the baseline feature set to train SVR models using an RBF kernel. SCATE-SVM2 additionally utilises an extra feature, which is based on the percentage of words that are labelled as BAD by the best word-level QE system (SCATE RF). External language resources from the IT domain are used to extract the additional features for both tasks.

UNBABEL (Task 2): Two systems were submitted for the word-level task. UNBABEL_2_linear is a feature-based linear sequential model. It uses the baseline features provided by the shared task organisers (with slight changes) conjoined with individual labels and pairs of consecutive labels. It also uses various syntactic dependency-based features (dependency relations, heads, and second-order structures like siblings and grandparents). The syntactic dependencies are predicted with `TurboParser` trained on the TIGER German treebank. UNBABEL_2_ensemble uses a stacked architecture, inspired by the last year's QUETCH+ system (Kreutzer et al., 2015), which combines three neural systems: one feedforward and two recurrent ones. The predictions of these systems are added as additional features in the linear system above. The following external resources were used: part-of-speech tags and extra syntactic dependency information obtained with `TurboTagger` and `TurboParser` (Martins et al., 2013), trained on the Penn Treebank (for English) and on the version of the German TIGER corpus used in the SPMRL shared task (Seddah et al., 2014) for German. For the neural models, pre-trained word embeddings from `Polyglot` (Al-Rfou et al., 2013) and those produced with a neural MT system (Bahdanau et al., 2014) were used.

USFD (Task 2): USFD's submissions tested two different approaches for phrase-level QE. The first one (CONTEXT submission) is an enhancement of the baseline feature set provided with the context features. The additional features consist of the source and target tokens which precede and follow the phrase under consideration, part-of-speech tags of these tokens, and language model scores for ngrams at the borders of the phrase. The second approach (W&SLP4PT submission) learns phrase-level labels from predictions at other levels. The models are trained on a set of seven features that are based on (i) the phrase segmentation itself (length and ratio to the sentence), (ii) word-level predictions (number of predicted OK/BAD words in the current phrase and in the sentence), and (iii) the predicted quality of the sentence. `CRFsuite` is used to train the prediction models in both cases.

USHEF (Task 3): Two different systems were submitted for Task 3. The first system (BASE-EMB-GP) combines the 17 baseline features with word embeddings from the source documents (English) using a Gaussian Process (GP) model. The word embeddings were learned by using the Continuous Bag-of-Words (CBOW) model (Mikolov et al., 2013), trained on the Google's billion-word corpus,[18] with a vocabulary size of 527K words. Document embeddings are extracted by averaging word embeddings in the document. The GP model was trained with two Rational Quadratic kernels (Rasmussen and Williams, 2006): one for the 17 baseline features and another for the 500 features from the embeddings. Since each kernel has its own set of hyperparameters, the full model can leverage the contributions from the two different sets. The second system (GRAPH-DISC) combines the baseline features with discourse-aware features. The discourse aware features are the same as the ones used by Scarton et al. (2015a) plus Latent Semantic Analysis (LSA) cohesion features (Scarton and Specia, 2014), number of subtrees and height of the Rhetorical Structure Theory (RST) tree and entity graph-based coherence scores (Sim Smith et al., 2016). Discourse-aware and RST tree features were extracted only for English (tools are only available for this language), LSA features were extracted for both languages, and entity graph-based coherence scores were extracted for the target language only (Spanish), as the source documents are expected to be coherent. This QE model was trained with an SVR algorithm.

UU (Task 1): The UU system uses SVR to predict HTER scores based on features extracted with `QuEst++` plus additional features. The feature vector consists of a combination of the 17 baseline features and top performing new features proposed by UU. These new features are related to reordering and noun translation, grammatical correspondence and structural integrity, based on parse trees and part-of-speech tags. The system submitted uses Kendall Tau distances in alignments between source and target for measuring reordering, noun group ratio, verb ratio and probabilistic context free grammars probabilities.

[18] `https://github.com/ciprian-chelba/1-billion-word-language-modeling-benchmark`

YSDA (Task 1): The YSDA submission is based on a simple idea that the more complex the sentence is the more difficult it is to translate. For this purpose, it uses information provided by syntactic parsing (information from parsing trees, some specific language constructions, etc). Additionally, it uses features based on pseudo-references, back-translation, web-scale language model, word alignments (as given by the data for Task 2), and combinations of several features. A regression model was training to predict BLEU as target metric instead HTER. The machine learning pipeline uses an SVR with RBF kernel to predict BLEU scores, followed by a linear SVR to predict HTER scores from BLEU scores. As external resources, the system uses a syntactic parser, pseudo-references and back-translation from web-scale MT system, and a web-scale language model.

6.3 Task 1: Predicting sentence-level quality

This task consists in scoring (and ranking) translation sentences according to the percentage of their words that need to be fixed. HTER (Snover et al., 2006) is used as quality score, i.e. the minimum edit distance between the machine translation and its manually post-edited version in [0,1].

As in previous years, two variants of the results could be submitted:

- **Scoring**: An absolute HTER score for each sentence translation, to be interpreted as an error metric: lower scores mean better translations.
- **Ranking**: A ranking of sentence translations for all source sentences from best to worst. For this variant, it does not matter how the ranking is produced (from HTER predictions or by other means). The reference ranking is defined based on the true HTER scores.

Data The data is the same as that used for the WMT16 Automatic Post-editing task, collected By the QT21 Project[19] in the Information Technology (IT) domain.[20] Source segments are English sentences and target segments are German translations produced by a strong SMT system built within the QT21 Project. The human post-editions

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

[20] The source sentences and reference translations were provided by TAUS (`https://www.taus.net/`) and come from a unique IT vendor.

are a manual revision of the target, done by professional translators using the PET post-editing tool (Aziz et al., 2012). HTER labels were computed using the TERCOM tool[21] with default settings (tokenised, case insensitive, exact matching only), and scores capped to 1.

As training and development data, we provided English-German datasets with 12,000 and 1,000 source sentences, their machine translations, post-editions and HTER scores. As test data, we provided an additional set of 2,000 English-German source-translations pairs produced by the same SMT system used for the training data.

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 and Spearman rho was computed using the William's test, following the approach suggested in (Graham, 2015).

Results Table 19 summarises the results for Task 1, 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 according to the ranking variant. We note that three systems have not submitted results ranking evaluation variant.

6.4 Task 2: Predicting word-level quality

The goal of this task is to evaluate the extent to which we can detect word-level errors in MT output. 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 OK and BAD tokens. The decision to bucket all error types together was made because of the lack of sufficient training data that could allow consideration of more fine-grained error tags.

Data This year's word-level task uses the same dataset as Task 1, for a single language pair: English-German. Each instance of the training, development and test sets consists of the following elements:

- Source sentence (English).
- Automatic translation (German).
- Manual post-edition of the automatic translation.
- Word-level binary (OK/BAD) labelling of the automatic translation.

The binary labels for the datasets were acquired automatically with the TERCOM tool. The tool identifies four types of errors: *substitution* of a word with another word, *deletion* of a word (word was omitted by the translation system), *insertion* of a word (a spurious word was added by the translation system), and word or sequence of words *shift* (word order error). Every word in the machine-translated sentence is tagged with one of these error types or not tagged if it matches a word from the reference.

All the untagged (correct) words were tagged with OK, while the words tagged with substitution and insertion errors were assigned the tag BAD. The deletion errors are not associated with any word in the automatic translation, so we could not consider them. We also disabled the shift errors by running TERCOM with the option '-d 0'. The reason for that is the fact that searching for shifts introduces significant noise in the annotation. The tool cannot discriminate between cases where a word was really shifted and where a word (especially common words such as prepositions, articles and pronouns) was deleted in one part of the sentence and then independently inserted in another part of this sentence, i.e. to correct an unrelated error. The statistics of the datasets are outlined in Table 20.

Evaluation This year's evaluation procedure is different from the one used in previous QE tasks. Previously, the submissions were evaluated in terms of F_1-score for the BAD class. However, this metric was criticised for being biased towards "pessimistic" labellings. It tends to rate higher the outputs of systems which labelled most of words as BAD, e.g. a trivial "all-BAD" baseline outperforms many real systems in terms of F_1-BAD score (Bojar et al., 2013).

Therefore, this year we used a different metric: the multiplication of F_1-scores of the BAD and OK classes (herein referred to as $\mathbf{F}_1$-**mult**). As it

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

System ID	**Pearson's r ↑**	MAE ↓	RMSE ↓	Spearman's ρ ↑	DeltaAvg ↑
English-German					
• YSDA/SNTX+BLEU+SVM	0.525	12.30	16.41	–	–
POSTECH/SENT-RNN-QV2	0.460	13.58	18.60	0.483	7.663
SHEF-LIUM/SVM-NN-emb-QuEst	0.451	12.88	17.03	0.474	8.129
POSTECH/SENT-RNN-QV3	0.447	13.52	18.38	0.466	7.527
SHEF-LIUM/SVM-NN-both-emb	0.430	12.97	17.33	0.452	7.886
UGENT-LT3/SCATE-SVM2	0.412	19.57	24.11	0.418	7.615
UFAL/MULTIVEC	0.377	13.60	17.64	0.410	7.114
RTM/RTM-FS-SVR	0.376	13.46	17.81	0.400	6.655
UU/UU-SVM	0.370	13.43	18.15	0.405	6.519
UGENT-LT3/SCATE-SVM1	0.363	20.01	24.63	0.375	7.008
RTM/RTM-SVR	0.358	13.59	18.06	0.384	6.379
BASELINE	0.351	13.53	18.39	0.390	6.300
SHEF/SimpleNets-SRC	0.320	13.92	18.23	–	–
SHEF/SimpleNets-TGT	0.283	14.35	18.22	–	–

Table 19: Official results for the scoring ad ranking variants of the WMT16 Quality Estimation Task 1. The systems are ranked according to the Pearson r metric and significance results are also computed for this metric. The winning submissions are indicated by a •. These are the top-scoring submission and those that are not significantly worse according to Williams test with 95% confidence intervals. The systems in the grey area are not different from the baseline system at a statistically significant level according to the same test.

	Sentences	Words	% of BAD words
Training	12,000	210,958	21.4
Development	1,000	19,487	19.54
Test	2,000	34,531	19.31

Table 20: Datasets for Task 2.

was shown in (Logacheva et al., 2016c), this metric is not biased neither towards "pessimistic" nor to "optimistic" labellings, and is good at discriminating between different systems.

We tested the significance of the results using randomisation tests (Yeh, 2000) with Bonferroni correction (Abdi, 2007).

Results The results for Task 2 are summarised in Table 21. We show the performance of all participating systems as well as the baseline model. The results are ordered by the F_1-mult metric. The top three submissions are statistically significantly different from any other system. However, we cannot unambiguously depict other significance groups in the table. Therefore, we only show the systems which are not significantly different from the baseline (grey area). The models above and below the grey area are significantly better and worse than the baseline system, respectively.

In order to show and analyse the groups of significantly different systems we plot the results of significance test as a heatmap (see Table 22). Here, a cell at the crossing of a row and a column corresponding to different submissions contains the information about the significance of the difference in their results: the darker the cell is, the lower is the significance in the difference for the pair of systems. The coloured frames denote groups of submissions which are not significantly different.

We should also note that in order to adequately evaluate the significance for multiple experiments we used Bonferroni correction. The essence of this method is that in cases when multiple results are compared (i.e. multiple comparisons are performed) the final significance level is computed as the initial significance level over the number of comparisons. In our case we had 91 comparisons which gave us $\alpha_B = \frac{\alpha}{91} = 0.0005$ for the significance level of 0.05. Bonferroni correction is quite a conservative method, so the number of significance groups may vary when using a different correction technique.

Overall, there are 10 groups of significantly different results: three of them contain one submission (the three best-performing models), other seven contain two to five models each (these are the groups denoted by frames of different colours).

6.5 Task 2p: predicting phrase-level quality

As an extension of the word-level task, we introduced a new task: phrase-level prediction. For this task, given a "phrase" (segmentation as given by the SMT decoder), participants are asked to label it as 'OK' or 'BAD'. Errors made by MT engines are interdependent and one incorrectly chosen word can cause more errors, especially in its local context. Phrases as produced by SMT decoders can be seen as a representation of this local context and in this task we ask participants to consider them as atomic units, using phrase-specific information to

System ID	F_1-**mult** ↑	F_1-BAD	F_1-OK
English-German			
• 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
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
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	0.324	0.368	0.880
RTM/s5-RTM-GLMd	0.308	0.349	0.882
UAlacant/SBI-Online	0.290	0.406	0.715
RTM/s4-RTM-GLMd	0.273	0.307	0.888

Table 21: Official results for the WMT16 Quality Estimation Task 2. The winning submissions are indicated by a •. These are the top-scoring submission and those that are not significantly worse according to approximate randomisation tests with 95% confidence intervals. The grey area indicates the submissions whose results are not statistically different from the baseline according to the same test.

improve upon the results of the word-level task.

Data The data to be used is exactly the same as for Task 1 and the word-level task. The labelling of this data was adapted from word-level labelling by assigning the 'BAD' tag to any phrase that contains at least one 'BAD' word. The phrase segmentation used in this dataset is the original segmentation of sentences produced by the SMT decoder during translation.

The dataset statistics are outlined in Table 23 (this is similar to Table 20, but shows the percentage of incorrect phrases instead of words).

Evaluation Although the QE was produced at the level of phrases, we used word-level metrics to evaluate the performance of participating systems. This choice was motivated by the fact that the length of phrases can vary significantly, and an incorrectly labelled phrase can actually mean 1 to 5 incorrectly labelled words, while phrase-level metrics do not weigh incorrect labels by the length of the phrases. We decided to use word-level evaluation to make the results of this task more intuitive. We used the same metric as the one used in task 2: multiplication of word-level F_1-OK and word-level F_1-BAD (F_1-mult). However, the test set was re-labelled in order to agree with phrase boundaries: if a phrase had at least one BAD word, all its labels were replaced with BAD.

Thus, the sequence

OK || BAD OK OK || OK || BAD OK || OK OK

was converted to:

OK || BAD BAD BAD || OK || BAD BAD || OK OK

As in Task 2, statistical significance was computed using randomisation tests with Bonferroni correction.

Results The results of the phrase-level task are represented in Table 24. Here, unlike the word-level task, we cannot find a single winner: although the F_1-mult scores of the top five systems vary from 0.379 to 0.364, this difference is not significant. However, all the winning submissions outperform the baseline.

Analogously to the previous task, we provide the F_1-BAD and F_1-OK scores in order to better understand the differences between the models. We can see that some models have very close F_1-mult scores, although their per class components scores can differ. For example, the F_1-mult scores of two submissions by the USFD team are very close (0.367 and 0.364). However, if we decompose these scores, we will see that both F_1-BAD and F_1-OK scores of the two models have around 2% of absolute difference: the W&SLP4PT model is more "pessimistic" (i.e. it is better at labelling BAD words), while the CONTEXT model identifies the correct words more accurately. However, the combinations of these scores lead to very similar F_1-mult. The situation is the same with all top five submissions: the differences in F_1-BAD are levelled off by the F_1-OK component, and the values of the F_1-mult are closer than those of F_1-BAD.

This suggests that the F_1-mult score might not be an best metric for the phrase-level task. While in the phrase-level models phrases of different length are treated in the same way, the word-level metric unfolds each phrase-level label to a set of

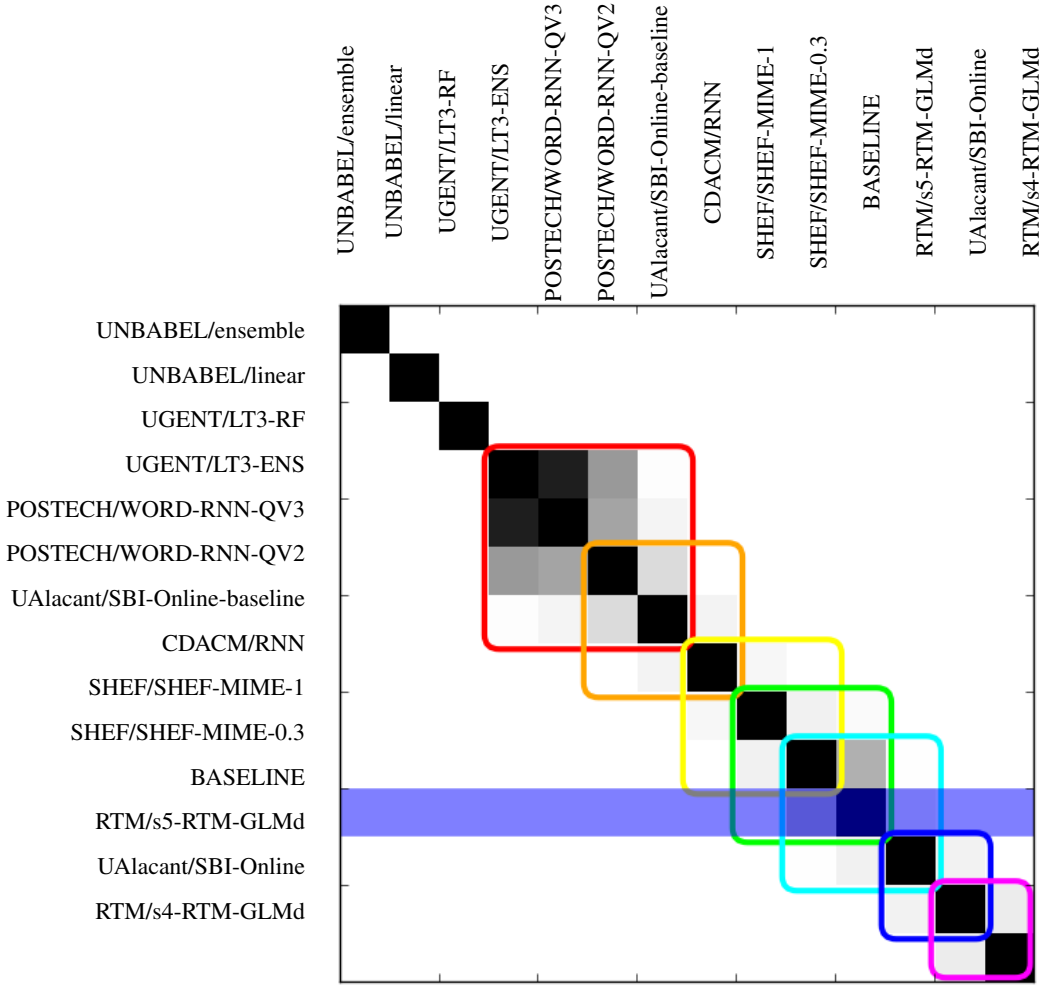

Table 22: Randomised significance test for the word-level task with Bonfferroni correction. The darker the cell, the lower the significance level of the difference between the scores of the corresponding systems. The coloured frames denote groups of submissions which are not significantly different. The blue row shows the baseline system.

	Sentences	Words	% of BAD words
Training	12,000	210,958	29.84
Development	1,000	19,487	30.21
Test	2,000	34,531	29.53

Table 23: Datasets for Task 2p.

word-level labels, thus giving different importance to phrases of different lengths. In order to find a more suitable metric we tested another evaluation strategy. We evaluated the submissions in terms of phrase-level F_1-scores: here all phrases were considered as uniform atomic units regardless of their lengths, and F_1-BAD and F_1-OK were computed as harmonic means of precision and recall for phrase-level of OK and BAD labels.

Table 25 shows the performance of phrase-level QE models measured in terms of multiplication of phrase-level F_1-scores. Except for some changes in the order of models, this ranking is very similar to the official one represented in Table 24. Here, the order of submissions by the POSTECH and CDACM teams is different from the ranking produced with the primary metric, but they are still not significantly different. On the other hand, the USFD team models are no longer best-performing under the phrase-level F_1-score. This evaluation shows that phrase-level F_1-mult is slightly better at discriminating between models, although they are still considered too close and no single best-performing approach can be identified.

6.6 Task 3: Predicting document-level quality

The document-level QE task consists in scoring and ranking documents according to their predicted quality. Knowing the quality of entire documents is useful for scenarios where fully automated approaches are used. An example is *gisting*, mainly if the user of the system does not know the source language. Another example are scenarios where post-editing is not an option or cannot be performed for the entire data.

Different from last year's task, in this second edition we use entire documents and a document-oriented quality score. The quality scores are achieved by a two-stage post-editing method (Scarton et al., 2015b), with post-editing done by professional translators. In the first stage, sentences are shuffled and post-edited without context (PE1). In the second stage, the post-edited sentences (from the first stage) are put together in the document context and post-edited again (PE2) by the same translator. This approach aims to isolate problems that can only be solved with document-level information.

Although the annotation task is considerably simple to perform, generating reliable quality labels from the data is not a trivial task. Average (AVG) and Standard Deviation (STDEV) of HTER between PE1 and MT ($PE_1 \times MT$), PE2 and MT ($PE_2 \times MT$) and PE2 and PE1 ($PE_2 \times PE_1$) are presented in Table 26.[22]

As shown in Table 26, $PE_1 \times MT$ and $PE_2 \times MT$ show low variation. As discussed last year (Bojar et al., 2015), we hypothesise that the low variation in the scores means that quality labels are not not able to distinguish documents reliably. $PE_2 \times PE_1$ values, on the other hand, show a high variation, indicating that the documents vary more when only document-wide errors are considered. However, taking only $PE_2 \times PE_1$ as quality label is not ideal as it disregards problems at word and sentence levels, which certainly also influence the quality of the document as whole. Our solution is to combine the scores such as to maintain a high enough variation in the data, while considering all issue levels. More specifically, we use a linear combination of $PE_1 \times MT$ and $PE_2 \times PE_1$ (Equation 1).

$$f = w_1 \cdot PE_1 \times MT + w_2 \cdot PE_2 \times PE_1, \qquad (1)$$

where w_1 and w_2 are empirically defined weights. w_1 was fixed to 1, while w_2 was optimised aiming at finding how much relevance we should give to each component in order to meet two criteria. First, the final label (f) should lead to significant data variation (in terms of standard deviation on the mean). Second, the difference between the MAE of the mean baseline[23] and the MAE of the official baseline QE system should be large enough.[24] The quality labels were defined by Equation 1 with $w_1 = 1$ and $w_2 = 13$.

[22] HTER was calculated by using the `Asiya` toolkit implementation of TER (non-tokenised and case insensitive) (Giménez and Màrquez, 2010).

[23] This baseline is calculated by assuming the mean of the training set as the predicted value of all instances in the test set.

[24] In our experiments, for variance we defined that the ratio between the standard deviation and mean should be at least 0.5 and for MAE difference, we defined it to be at least 0.1. w_2 was increased by 1 at each iteration and the optimisation process stopped when any of the requirements was met.

System ID	F_1-**mult** ↑	F_1-BAD	F_1-OK
English-German			
• CDACM/RNN	0.380	0.503	0.755
• POSTECH/PHR-RNN-QV3	0.378	0.495	0.764
• POSTECH/PHR-RNN-QV2	0.369	0.478	0.772
• USFD2/W&SLP4PT	0.368	0.486	0.757
• USFD2/CONTEXT	0.365	0.470	0.777
RTM/s5_RTM-GLMd	0.327	0.408	0.802
BASELINE	0.321	0.401	0.800
RTM/s4_RTM-GLMd	0.307	0.377	0.814
Ualacant/SBI-Online-baseline	0.259	0.493	0.526
UAlacant/SBI-Online	0.098	0.459	0.213

Table 24: Official results for the WMT16 Quality Estimation Task 2p. The winning submissions are indicated by a •. These are the top-scoring submission and those that are not significantly worse according to approximate randomisation tests with 95% confidence intervals. The grey area indicates the submissions whose results are not statistically different from the baseline.

System ID	F_1-**mult** ↑	F_1-BAD	F_1-OK
English-German			
• POSTECH/PHR-RNN-QV3	0.393	0.518	0.759
• POSTECH/PHR-RNN-QV2	0.388	0.504	0.771
• CDACM/RNN	0.378	0.500	0.756
USFD/CONTEXT	0.364	0.467	0.780
USFD/W&SLP4PT	0.363	0.475	0.764
RTM/s5-RTM-GLMd	0.331	0.413	0.802
BASELINE	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
UAlacant/SBI-Online	0.146	0.456	0.320

Table 25: Results for the WMT16 Quality Estimation Task 2p computed in terms of phrase-level F_1-scores. The winning submissions are indicated by a •. These are the top-scoring submission and those that are not significantly worse according to approximate randomisation tests with 95% confidence intervals. The grey area indicates the submissions whose results are not statistically different from the baseline.

	$PE_1 \times MT$	$PE_2 \times MT$	$PE_2 \times PE_1$
AVG	0.346	0.381	0.042
STDEV	0.108	0.091	0.034
Ratio	0.312	0.239	0.810

Table 26: AVG and STDEV of the post-edited data.

Data The documents were extracted from the WMT translation task test data from 2008 to 2013, using submissions from all participating MT systems. Source documents were randomly chosen. For each source document, a translation was taken from a different MT system. We considered EN-ES as language pair, extracting 208 documents. All documents were post-edited as previously explained. 146 documents were used for training and 62 for test.

Evaluation The evaluation of the document-level task was the same as that for the sentence-level task. Pearson's r, MAE and RMSE are reported as evaluation metrics for the scoring task, with Pearson's r as official metric for the ranking of systems. For the ranking task, Spearman's ρ correlation and DeltaAvg are reported, with Spearman's rho as main metric. The significance of the results is evaluated by applying the Williams test on Pearson's r scores.

Results The results of both the scoring and ranking variants of the task are given in Table 27, sorted from best to worst by using the Pearson's r scores as primary key. USHEF/BASE-EMB-GP and RTM/RTM-FS+PLS-TREE showed the best scores, with no significant difference between them. The other two systems are not statistically significantly different from the baseline.

The two winning submissions are very different. The BASE-EMB-GP system combines word embeddings with the official baseline features in a GP model with two-kernels, while RTM-FS+PLS-TREE is an RTM implementation that explores more sophisticated features from the source and target texts. For ranking variant, however, RTM-FS+PLS-TREE showed better results. Moreover, this is the only system with higher scores than the baseline that is also significantly better than the baseline.

6.7 Discussion

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

System ID	**Pearson's** $r \uparrow$	MAE $\downarrow$	RMSE $\downarrow$	Spearman's $\rho \uparrow$	DeltaAvg $\uparrow$
English-Spanish					
• USHEF/BASE-EMB-GP	0.391	0.295	0.128	0.393	0.111
• RTM/RTM-FS+PLS-TREE	0.356	0.253	0.118	0.476	0.123
RTM/RTM-FS-SVR	0.293	0.268	0.125	0.360	0.119
BASELINE	0.286	0.278	0.139	0.354	0.093
USHEF/GRAPH-DISC	0.256	0.285	0.144	0.285	0.061

Table 27: Official results for the scoring ad ranking variants of the WMT16 Quality Estimation Task 3. The systems are ranked according to the Pearson r metric and significance results are also computed for this metric. The winning submissions are indicated by a •. These are the top-scoring submission and those that are not significantly worse according to Williams test with 95% confidence intervals. The systems in the grey area are not different from the baseline system at a statistically significant level according to the same test.

previously identified for it.

Domain specific, professionally done post-editions

Last year we used the largest dataset of all editions of the shared task to date (for sentence and phrase-level QE): ∼14K segment pairs altogether. However, the findings were somewhat inconclusive as the quality of the dataset was dubious (crowd-sourced post-editions). This year we were able to collect a dataset of comparable size (15K) but in a completely controlled way, and with professional (paid) translators to ensure the quality of the data. Another critical difference in this year's main dataset is its domain: IT, as opposed to the rather general, "news" domain that had been used so far. Finally, we had access to the SMT system that produced the translations, which was very important for the new task introduced this year – phrase-level QE. For phrase-level QE, the segmentation of the sentences in phrases was necessary. Having a more repetitive text domain was deemed particularly relevant for the word and phrase-level tasks, where data sparsity is a major issue.

In practice, we found that this year's main dataset is similar to last year's in terms of error distribution at the word-level: about 20% of the words are labelled as BAD. One thing to notice, however, is that with the new data systems did not seem to benefit from filtering data out. Last year, various systems reported improvements from filtering out significant portions of the "all/mostly GOOD" sentences, which could have meant that these sentences may not have been correct, but did not get post-edited by the crowdworkers.

In terms of progress with respect to last year for comparable tasks, although direct comparisons are not possible, we observed that:

- For sentence-level, the Pearson correlation of the winning submission last year was 0.39 (against 0.14 of the baseline system). This year, the winning submission reached 0.52 Pearson correlation, with many other systems above 0.4 (against 0.35 of the same baseline system as last year). One can speculate that the task was made somewhat "easier" by using high quality data, but the delta in Pearson correlation between the baseline and winning submission is still very substantial.

- For word-level, the main metric used this year (F_1-mult) is different from the one used last year (F_1-BAD), and this may have been the metric most systems optimised against, so looking at the F_1-BAD results for both years is not entirely fair to this year's systems, but nevertheless this year's systems performed much better: 0.56 against 0.43 last year. The baseline system used last year was much simpler, and therefore comparisons against the baseline cannot be made.

Effectiveness of new quality label provided by humans for document-level prediction

Participation in the document-level task was again disappointingly low, with only four systems. Document-level QE is still a relative new area and engaging the community is therefore still a challenge.

The main changes in this year's task were the fact that entire documents are used (potentially resulting in the need for more discourse/document-wide features), and the the fact that the quality labels are computed based on human post-editing. We start by analysing the new quality label against automatic metrics (such as BLEU) used in previous work. Our hypothesis is that automatic metrics are not reliable labels for document-level evaluation (as discussed in (Scarton et al., 2015b)). Therefore, we expect that our new label would perform differently from these metrics. We use cor-

relation to measure whether or not the new label shows different behaviour. Table 28 shows Pearson r correlation scores for automatic metrics versus the new label, as well as between HTER and all labels. The HTER score was calculated considering the last version of the two-stage post-editing method ($PE_2 \times MT$).

	NEW (↓)	BLEU(↑)	TER(↓)	METEOR(↑)
BLEU	-0.168	-	-	-
TER	$+0.195$	-0.928	-	-
METEOR	-0.186	$+0.954$	-0.961	-
HTER(↓)	**+0.516**	-0.462	$+0.449$	-0.452

Table 28: Pearson r correlation between automatic metrics, our new label (NEW) and HTER. All correlation scores are significant with 95% of confidence.

Although the new label showed some correlation to BLEU, TER and METEOR, the best correlation is showed with HTER. On the other hand, the automatic metrics showed higher correlation among themselves than against HTER scores, which is expected since such metrics are similar in many ways.

An important observation is that the automatic metrics are calculated against a human translation and HTER is calculated against a post-edited version. The effect of this is that BLEU, TER and METEOR compare the MT output to a human translation that can be completely different from the MT output, without necessarily meaning that the machine translation is bad. HTER, conversely, compares the MT output to its post-edited version.

It is also worth noticing that although HTER did not show a high variation (0.091 for mean 0.381 - third column of Table 26), similar to the automatic metrics, it still did not show very high correlation with BLEU, TER and METEOR. Conversely, the new label showed high correlation with HTER, but much lower correlation with BLEU, TER and METEOR than HTER itself. This seems to indicate that the new label captures different information than BLEU, TER and METEOR. Therefore, we believe that the new label and standard evaluation metrics provide complementary information on translation quality.

In terms of features, most are similar to those used by the systems submitted last year, which are aggregations of sentence-level feature values. Therefore, our hypothesis that discourse/document-aware features would show better results on evaluating full document was not proved. Systems using discourse-aware features (USHEF/GRAPH-DISC) did not show improvements relative to the baseline system. This could be an indication of the limitations of the features or of the labels themselves.

QE at the phrase level

One of the main motivations for switching from the word level to phrase level is the fact that MT errors are often context-dependent, and the wrong choice of a word might be explained by an error in its context. A good example of such errors are adjectives that take the gender of the noun they depend on, and become erroneous if this noun is replaced with another noun of a different gender.

This motivation suggests that the phrases to be used as atomic units in a phrase-level QE system should be syntactically motivated. However, there can be other approaches. For example, the very popular SMT systems manipulate sequences of words as opposed to single words. These sequences – referred to as "phrases" – are not linguistically motivated phrases. During decoding these phrases are selected or rejected as atomic units (regardless of the quality of the individual words they consist of), and thus it may be useful to estimate the quality of the entire phrase.

Overall, there is no single answer to what should be considered as a "phrase" in a phrase-level QE system. A fully-fledged phrase-level QE system should be able to handle both the segmentation of a sentence into phrases and the labelling of each phrase for quality. However, each of these two steps is a complex problem on itself. Therefore, for the first edition of the task we decided to simplify it and provide the phrase segmentation. Following Logacheva et al. (2015), we considered a "phrase" the final segmentation produced by the SMT decoder by an MT decoder that generated the automatic translations in the dataset. This segmentation is useful for decoding-time QE.

The baseline phrase-level QE system uses a set of features which were originally designed for sentences and later adapted for smaller sequences. These features were used to train a CRF model. Participants chose many different techniques to model the task. The best performing ones are deep neural networks: the Recurrent Neural Network from the POSTECH team which predicts the phrase-level label and the CDACM Recurrent Neural Network whose word-level predictions were successfully applied to the phrase-level task. Two of the submitted models make use of the baseline feature set: the USFD team enhanced

it with context information, while the UAlacante team combined it with features based on pseudo-reference translations coming from a number of sources.

Several teams attempted to take into account the predictions for other the task at other levels. The phrase-level submission from CDACM simply labels the phrase-level test set using word-level predictions; while the UAlacant submission uses the probability of each word in a phrase being labelled as BAD along with other external features. Similarly, USFD uses information on word labels within a phrase as well as the information on sentence-level quality.

Comparison of word-level and phrase-level models The word-level and phrase-level systems that participated in Tasks 2 and 2p are not directly comparable. Although they are evaluated on the same test sentences, and the labels for the test set come from the same post-editions, they are not identical. The labels for the phrase-level test set were modified in order to comply with the phrase-level training data. We established a pessimistic approache where a phrase is considered BAD if any of its words is BAD. We changed the word-level labels so that all labels within a BAD phrase are also BAD. This is analogous to replacing some OK labels with BAD labels for words.

Nevertheless, we can still try to compare the word-level and phrase-level submissions if we change the word-level submissions appropriately. Let us consider that a word-level QE model was used to label phrases for quality. Following the rules mentioned above we will label a phrase as BAD if our QE model labelled any of words of this phrase as BAD. After performing this transformation we can use the Task 2p test set to evaluate both phrase-level and (modified) word-level submissions.

While this comparison is an approximation as the submitted word-level models were not trained to predict the quality of phrases, it still allows a rough comparison between word-level and phrase-level QE models. One of the purposes of the phrase-level task was to understand if the subsentence-level QE can benefit from joint labelling of groups of words, and this cross-task comparison is a means to try to answer that question.

Table 29 contains the joint results of Tasks 2 and 2p. The best-performing system is the winning word-level submission. Moreover, the word-level systems tend to perform better in this task in general: the top seven positions in this joint table are occupied by the word-level systems. Some of the phrase-level systems which performed well turn out not to be better than the word-level baseline system. Presumably, this result means that defining the quality for individual words yields better results in general.

Another observation we can make from this table is the change in the significance level of the results: some of the word-level submissions which were significantly different from the word-level baseline model in the original (word-level) task are no longer different in the phrase-level version. This can shed some light on the difficulties we had with defining the single best phrase-level system: perhaps the lack of significance in the differences between the labellings is derived from the phrase-level task itself. Alternatively, as it was discussed in Section 6.5, it could be explained by the fact that F_1-mult score is not a suitable metric for phrase-level QE.

In order to examine how the phrase-level task relates to the word-level one more closely we performed a different comparison. Some of the teams presented their results for both variants of Task 2, and the majority of them have similar models for both levels: they tried to adapt their original word-level system for the phrase-level task. We can compare these pairs of systems to see if the adaptation was successful. This is not a direct comparison, because the models, although similar, cannot be identical due to differences between words and phrases. This comparison was only done for analysis, as it can give us more insights on the future perspectives for the phrase-level task. Table 30 outlines the results of this comparison.[25]

Here, in order to enable the direct comparison, we adapted the word-level systems to phrase-level test set the same way as we did for Table 29. It can be clearly seen that the performance of word-level systems is better than that of the analogous phrase-level systems. There are multiple possible reasons for that, for example, wrong choice of phrase-level features, limitations of models originally designed for word-level QE in dealing effectively with word

[25]The submission by the CDACM team was not included in the table because their phrase-level submission is an adaptation of word-level predictions to phrase level. It was performed analogously to our word-level submissions adaptation, therefore it should be no different.

	System ID	F_1-**mult** ↑
	English-German	
• **word**	UNBABEL/ensemble	0.517
word	UNBABEL/linear	0.487
word	UGENT-LT3/SCATE-RF	0.426
word	POSTECH/WORD-RNN-QV3	0.399
word	UGENT-LT3/SCATE-ENS	0.395
word	POSTECH/WORD-RNN-QV2	0.388
word	CDACM/RNN	0.381
phrase	CDACM/RNN	0.379
phrase	POSTECH/PHR-RNN-QV3	0.378
phrase	POSTECH/PHR-RNN-QV2	0.369
word	UAlacant/SBI-Online-baseline	0.369
phrase	USFD/W&SLP4PT	0.367
word	SHEF/SHEF-MIME-0.3	0.367
word	SHEF/SHEF-MIME-1	0.367
phrase	USFD/CONTEXT	0.364
word	BASELINE	0.360
word	RTM/s5-RTM-GLMd	0.344
phrase	RTM/s5-RTM-GLMd	0.327
phrase	BASELINE	0.321
word	RTM/s4-RTM-GLMd	0.313
phrase	RTM/s4-RTM-GLMd	0.307
word	UAlacant/SBI-Online	0.290
phrase	UAlacant/SBI-Online-baseline	0.259
phrase	UAlacant/SBI-Online	0.097

Table 29: Comparison of submissions for Tasks 2 and 2p in terms of word-level F_1-mult scores computed on the test set used for the Task 2p. Word-level systems (Task 2) are indicated by "**word**", while phrase-level systems (Task 2p), by "**phrase**". The winning submission is indicated with •. The grey area indicates the models which are not significantly different from the word-level baseline system, the cyan area indicates the models which are not significantly different from the phrase-level baseline.

System ID	Word-level	Phrase-level
English-German		
POSTECH/RNN-QV3	0.399	0.378
POSTECH/RNN-QV2	0.388	0.369
RTM/s5-RTM-GLMd	0.344	0.327
RTM/s4-RTM-GLMd	0.313	0.307
Ualacant/SBI-Online-baseline	0.369	0.259
Ualacant/SBI-Online	0.290	0.097

Table 30: Comparison of systems' performance in Task 2 (word-level) and 2p (phrase-level). Performance is evaluated in terms of word-level F_1-mult scores computed on the test set used for the Task 2p. The submissions to the word-level task are modified in order to comply with the phrase-level task.

sequences.

Nevertheless, it is worth noticing the phrase-level QE systems introduced a number of interesting strategies that allowed them to outperform a strong baseline phrase-level model. Finally, we recall that the evaluation metric – word-level F_1-mult – has difficulties to distinguish phrase-level systems. This suggests that we may need to find a different metric for evaluation of the phrase-level task, with phrase-level F_1-mult one of the candidates.

7 Automatic Post-editing Task

This year WMT hosted the second round of the shared task on MT automatic post-editing (APE), which consists in automatically correcting the errors present 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.

Also this year, the general framework consisted in a "black box" scenario in which the MT system that produced the translations is unknown to the participants and cannot be modified. However, building on the lessons learned in the first pilot round (Bojar et al., 2015), some changes have been made.

The major differences concern the domain and the origin of the data. First, we moved from the general `news` domain to the more specific `information technology` (IT) domain. This novelty is motivated by the difficulties observed in the pilot round, in which the baseline (the simple *do-nothing* APE system that leaves all the test sentences unmodified) remained unbeaten. Indeed, the scarce repetitiveness of the `news` domain prevented participants to learn from the training data effective correction patterns that are also applicable to the test set. Second, concerning the origin of the data, we moved from post-edits obtained from non-professional crowdsourced workforce to material collected from professional translators. Data collected from trained professionals represents first of all a more standard scenario for the translation industry. Besides this, they are considered to guarantee higher translation coherence, feature higher repetitiveness and, eventually, make the APE task more feasible by automatic systems.

Other changes concern the language combination and the evaluation mode. As regards the languages, we moved from English-Spanish to English-German, which is one of the language pairs covered by the QT21 Project[26] that supported data collection and post-editing. Concerning the evaluation, we changed from TER scores computed both in case-sensitive and case-insensitive mode to a single ranking based on case sensitive measurements.

Besides these changes the new round of the APE task included some extensions in the evaluation. BLEU (Papineni et al., 2002) has been introduced as a secondary evaluation metric to measure the improvements over the rough MT output. In addition, to gain further insights on final output quality, a subset of the outputs of the submitted systems has also been manually evaluated.

Based on these changes and extensions, the goals of this year's shared task were to: *i)* improve and stabilize the evaluation framework in view of future rounds, *ii)* analyze the effect on task feasibility of data coming from a narrow domain, *iii)* analyze the effect of post-edits collected from professional translators, *iv)* analyze how humans perceive TER/BLEU performance differences between different systems, *v)* measure the progress made during one year of research on the APE task.

Although the changes made with respect to the first pilot round prevent from fair and informative result comparisons, we believe that these objectives were successfully achieved. Most noticeably, the higher feasibility of the task brought by domain-specific data and professional post-edits resulted in significant baseline improvements (up to 3.2 TER and 5.5 BLEU points), which are also evident to human evaluation. These positive results, together with the increase in the number of participants with respect to the pilot round (from four to six), represent a good starting point for future rounds of the APE task.

7.1 Task description

Similar to last year, 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.

7.1.1 Data

One of the findings of the first pilot task was that the origin and the domain of the data pose specific challenges to the participating systems. In particular, our analysis highlighted the strong dependence of system results on data repetitiveness, which tends to be higher within restricted domains and with coherent post-edits. On one side, restricted domains are more likely to feature smaller vocabularies and to be more repetitive (or, in other terms, less sparse). This situation, in turn, will likely determine a higher applicability of the learned error correction patterns. On the other side, coherent post-edits (like those produced within controlled professional environments) will result in a lower variability in the correction of specific errors and, in turn, in favorable conditions to learn and gather reliable statistics. These considerations motivate some of the major changes of this year's round of the APE task, namely those concerning the domain (a specific one as opposed to `news`) and the origin of the post-edits (from professional translators instead of crowdsourced).

The data used this year was released by the QT21 Project. This material was obtained by

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

randomly sampling from a collection of English-German (*source*, *target*, *human post-edit*) triplets drawn from the `Information Technology` (IT) domain.[27] Also this year, the main reason for random sampling was to induce a higher data homogeneity and, in turn, to increase the chances that correction patterns learned from the training set can be applied also to the test set. The downside of losing information yielded by text coherence (an aspect that some APE systems might take into consideration) has hence been accepted in exchange for a higher error repetitiveness across the three data sets.

The training and development sets respectively consist of $12,000$ and $1,000$ instances. In each instance:

- The source (SRC) is a tokenized English sentence whose length ranges between 3 and 30 tokens;
- The target (TGT) is a tokenized German translation of the source. Translations were obtained with a statistical MT system.[28] 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.[29]

Test data ($2,000$ instances) 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.

Table 31 provides some basic statistics about the data. As discussed in Section 7.3, the differences in the domain and the origin of this year's data can contribute to explain the large improvements over the baseline, which in the first pilot round unfortunately remained unbeaten. These differences are highlighted by the Repetition Rate (RR[30]) scores reported in Table 32. Values are indeed very close to those observed in the IT-related corpus (the Autodesk Post-Editing Data corpus[31]) that was used last year as a term of comparison to motivate the high difficulty of dealing with `news` data.

7.1.2 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). Differently from the first edition of the task, in which this distance was only measured in terms of Translation Error Rate (TER) (Snover et al., 2006), this year the BLEU (Papineni et al., 2002) score was also used. TER is an evaluation metric commonly used in MT-related tasks (e.g. in quality estimation) to measure the minimum edit distance between an automatic translation and a reference translation.[32] BLEU is the reference metric for MT evaluation and is based on modified n-gram precision to find how many of the n-grams in the candidate translation are present in the reference translation over the entire test set. The main difference between the two metrics is that TER works at word level, while BLEU takes advantage of words and n-grams with *n* from 2 to 4. Systems were ranked based on the average TER calculated on the test set by using the TERcom[33] software: lower average TER scores correspond to higher ranks. BLEU was computed using the multi-bleu.perl package[34] available in MOSES.

Differently from the pilot round, in which TER was computed both in case-sensitive and case-insensitive mode, this year we opted for only one mode. Working with German, for which case errors are of crucial importance, participants' submissions were evaluated with the more strict case-sensitive mode.

[27] The source sentences (together with their reference translations which were not used for the task) were provided by TAUS (`https://www.taus.net/`) and originally come from a unique IT vendor.

[28] It consists of a phrase-based machine translation system leveraging generic and in-domain parallel training data and using a pre-reordering technique (Herrmann et al., 2013). It takes also advantages of POS and word class-based language models.

[29] German native speakers working at Text&Form `https://www.textform.com/`.

[30] Repetition rate measures the repetitiveness inside a text by looking at the rate of non-singleton n-gram types (n=1...4) and combining them using the geometric mean. Larger value means more repetitions in the text.

[31] `https://autodesk.app.box.com/Autodesk-PostEditing`

[32] Edit distance is calculated as the number of edits (word insertions, deletions, substitutions, and shifts) divided by the number of words in the reference. Lower TER values indicate lower distance from the reference as a proxy for higher MT quality.

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

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

	Tokens			Types			Lemmas		
	SRC	TGT	PE	SRC	TGT	PE	SRC	TGT	PE
Train (12,000)	201,505	210,573	214,720	9,328	14,185	16,388	5,628	11,418	13,244
Dev (1,000)	17,827	19,355	19,763	2,931	3,333	3,506	1,922	2,686	2,806
Test (2,000)	31,477	34,332	35,276	3,908	4,695	5,047	2,479	3,753	4,050

Table 31: Data statistics.

	APE@WMT15 (EN-ES, news, crowd)	APE@WMT16 (EN-DE, IT, prof.)
SRC	2.905	6.616
TGT	3.312	8.845
PE	3.085	8.245

Table 32: Repetition Rate (RR) of the WMT15 (English-Spanish, `news` domain, crowdsourced post-edits) and WMT16 (English-German, `IT` domain, professional post-editors) APE Task data.

7.1.3 Baseline

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 system that leaves all the test targets unmodified.[35] Baseline results are reported in Table 34.

Monolingual translation as another term of comparison. To get some insights about the progress with respect to the first pilot task, participating systems were also evaluated against a re-implementation of the approach firstly proposed by Simard et al. (2007).[36] Last year, in fact, this statistical post-editing approach represented the common backbone of all submissions (this is also reflected by the close results achieved by participants in the pilot task). For this purpose, a phrase-based SMT system based on Moses (Koehn et al., 2007) was used. 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 APE 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 Table 34.

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

7.2 Participants

This year, six teams (two more than in the pilot round) participated in the APE task by submitting a total of eleven runs. Participants are listed in Table 33; a short description of their systems is provided in the following.

Adam Mickiewicz University. This system is among the very first ones exploring the application of neural translation models to the APE task. In particular, it investigates the following aspects: *i)* the use of artificially-created post-edited data to train the neural models, *ii)* the log-linear combination of monolingual and bilingual models in an ensemble-like manner, *iii)* the addition of task-specific features in the log-linear model to control the final output quality. Concerning the data, in addition to the official training and development material, the system exploits the English-German bilingual training material released for the IT-domain and news translation shared tasks. The German monolingual common crawl corpus admissible for these two tasks is also exploited. This data is used by a "round-trip translation" approach aimed to artificially create the huge amount of triples needed to train the neural models. Such models are attentional encoder-decoder models (Bahdanau et al., 2014) trained with subword units (Sennrich et al., 2015) in order to deal with the limited ability of neural translation models to handle out-of-vocabulary words. They include both monolingual models trained to translate from TGT to PE, and cross-lingual models trained to translate from SRC to PE. An ensemble is obtained through their log-linear combination with empirically-set weights (higher for the

[35] 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.

[36] This is done based on the description provided in (Simard et al., 2007). Our re-implementation, however, is not meant to officially represent such approach. Discrepancies with the actual method are indeed possible due to our misinterpretation or to wrong guesses about details that are missing in the paper.

ID	Participating team
AMU	Adam Mickiewicz University, Poland (Junczys-Dowmunt and Grundkiewicz, 2016)
CUNI	Univerzita Karlova v Praze, Czech Republic (Libovický et al., 2016)
DCU	Dublin City University, Ireland
FBK	Fondazione Bruno Kessler, Italy (Chatterjee et al., 2016)
JUSAAR	Jadavpur University, India & Saarland University, Germany
USAAR	Saarland University, Germany (Pal et al., 2016)

Table 33: Participants in the WMT16 Automatic Post-editing task.

TGT-to-PE model). Finally, a task-specific feature based on string matching is added to the log-linear combination to control the faithfulness of the APE results with regard to the input. This is done by penalizing words in the output that do not appear in the input to be corrected.

Univerzita Karlova v Praze. Also this system is based on the neural translation model with attention proposed by Bahdanau et al. (2014) and extends it to include multiple encoders able to manage different input representations. Each encoder is a bidirectional RNN that takes in input a one-hot vector for each representation of a word. The decoder is an RNN which receives an embedding of the previously produced word as an input in every time step together with the hidden state from the previous time step. The RNNs output is then used to compute the attention and the next word distribution. The attention is computed over each of the encoders separately. The initial state of the decoder is obtained by a weighted combination of the encoders final states. To improve the capability of the network to focus on the edits made by the post-editors, the target sentence is converted in the minimum-length sequence of edit operations performed on the machine-translated sentence. For this purpose, the network vocabulary is extended adding two more tokens (keep and delete) and the new representation is made of a sequence of keep, delete and insert operations, where the insert operation is defined by placing the word itself. The different inputs used for the APE task submission are the source sentence and its translation into the target language and the sequence of edits. The network is trained using only the task data. To better handle the complexity of the German target language, different language-dependent pre- and post-processing are used, in particular, splitting the contracted prepositions and articles and separating some pronouns from their case ending.

Dublin City University. This system is designed as an automatic rule learning system. It considers four types of editings, i.e. replacement, deletion, insertion and reordering, as generalized replacement (GR) editings. GR editings are learned from aligning words in source and target sentences and records replacement pairs and their corresponding contexts for each source and target sentence pair. When the source word is empty, it is of an insertion editing; similarly, when the target word is empty, it is of a deletion editing. When the source words and target words in a GR editing both comprise the same set of words but with different orderings, it is of a reordering editing. The word-based GR editings and their generalization which uses POSs to replace their context words, comprise the whole rule set of GR editings. There is no linguistic knowledge incorporated in the system, which therefore can be applied to any language for post-editing purposes. Three things are learned from the training set, 1) the GR rules, 2) the precedence ordering of these rules, and 3) the maximum number of rules to be applied to a sentence. For each set of GR rules, the precedence ordering can be ranked based on the counts of replacement words, the counts of their context words, the lengths of GR editings, the number of occurrences of GR editings observed in training set and/or their combinations. In the training phase, given a set of GR rules, the system will apply the rules to the training set using different settings of precedence ordering and maximum number of rules to be applied for each sentence. The system is trained when one setting is selected if the system yields the best overall post-edited results by applying that setting. In the test phase, the GR rules will be applied to each sentence in the test set using the trained precedence ordering and stop when the maximum number of rules to be applied is met for that sentence.

Fondazione Bruno Kessler. This system combines the monolingual statistical approaches previously exploited in Chatterjee et al. (2015a) with a factored machine translation model that is able to leverage benefits from both. One is the monolingual statistical translation approach proposed by Simard et al. (2007). The other is the context-aware variant proposed by Béchara et al. (2011). The former is more robust and it better generalizes the learned post-editing rules. The latter is prone to data sparsity, word alignment and tuning problems due to its richer representation of the terms. Nevertheless, by integrating knowledge about the source context in the learned rules, its precision is a good complement to the higher recall of (Simard et al., 2007). By enabling a straightforward integration of additional annotation (factors) at the word-level, factored translation models (Koehn and Hoang, 2007) are used to leverage such complementarity. In the FBK system they include part-of-speech-tag and class-based neural language models (LM) along with statistical word-based LM to improve the fluency of the post-edits. These models are built upon a data augmentation technique (i.e. the extension of the monolingual parallel corpus with the post-edits available in the training data), which helps to mitigate the problem of over-correction in phrase-based APE systems. One of the submitted runs incorporates a quality estimation model (C. de Souza et al., 2013, 2014), which aims to select the best translation between the MT output and the automatic post-edit.

Jadavpur University & Saarland University. This system contains three basic components: statistical APE, word deletion model and word surface form correction model. The final generated translation is the product of a multi-engine re-ranking system. The statistical APE component is based on the phrase-based APE approach of Pal et al. (2015). MT outputs generally contain four types of errors: presence of unwarranted words, wrong word surface form, absence of some relevant words, and wrong word order. The system tries to address the first two types of errors. The word deletion model is based on source language context modelling and target language word deletion frequency in the training data. The surface form correction model tries to fix the morphological errors by generating all possible surface forms for each root word present in the MT output and to select the most likely sequence of word surface forms by applying a language model. The word deletion model and the word surface form correction model are applied to all the APE outputs. Finally, the generated translation candidates are ranked using a ranking algorithm based on language model information and a length-based heuristic. The top ranked output is chosen as the final APE output.

Saarland University. This system combines the Operation Sequence Model (OSM) (Durrani et al., 2011) with the classic phrase-based statistical MT (PB-SMT) approach. The OSM-APE method represents the post-edited translation process as a linear sequence of operations such as lexical generation of post-edited translation and their orderings. The translation and reordering decisions are conditioned on n previous translation and reordering decisions. This technique is able to model both local and long-range reorderings that are quite useful when dealing with the German language. To improve the capability of choosing the correct edit to process, eight new features are added to the log-linear model. These features capture the cost of deleting a phrase and different information on possible gaps in reordering operations. The monolingual alignments between the MT outputs and their post-edits are computed using different methods based on TER, METEOR (Snover et al., 2006) and Berkeley Aligner (Liang et al., 2006). Only the task data is used for these submissions.

7.3 TER/BLEU results

The official TER and BLEU results achieved by participants are reported in Table 34. The submitted runs are sorted based on the average (case-sensitive) TER measured on test data, which was this year's primary evaluation metric.

Looking at the performance of the two baselines, i.e. the raw MT output (Baseline) and the basic statistical APE approach of Simard et al. (2007), the latter outperforms the former with both metrics. This indicates that, under this year's evaluation conditions, the MT outputs could be improved by learning from human post-editors' work.

Differently from the pilot task (Bojar et al., 2015), in which none of the runs was able to beat the baselines, this year half of the participants achieved this goal by producing automatic post-edited sentences that result in lower TER (with a

ID	Avg. TER	BLEU
AMU Primary	21.52	67.65
AMU Contrastive	23.06	66.09
FBK Contrastive	23.92	64.75
FBK Primary	23.94	64.75
USAAR Primary	24.14	64.10
USAAR Constrastive	24.14	64.00
CUNI Primary	24.31	63.32
(Simard et al., 2007)	24.64	63.47
Baseline	24.76	62.11
DCU Contrastive	26.79	58.60
JUSAAR Primary	26.92	59.44
JUSAAR Contrastive	26.97	59.18
DCU Primary	28.97	55.19

Table 34: Official results for the WMT16 Automatic Post-editing task – average TER (↓), BLEU score (↑).

maximum of -3.24 points) and higher BLEU score (up to +5.54 points). All differences with respect to such baselines are statistically significant. This suggests that the correction patterns learned from the data were reliable enough to allow most systems to effectively correct the original MT output.

The obvious question is whether the improvements observed this year are due to the new data set (i.e. domain-specific texts and professional post-edits) or to a real technology jump (i.e. the use of neural end-to-end APE systems, factored or operational sequential models). A partial answer is given by the performance of the approach of Simard et al. (2007), which we run on the data of both rounds of the APE task with the same implementation. Although its results on the two test sets are difficult to compare (also due to the different language setting), the overall TER scores and the relative distances with respect to the other submitted runs can give us some indications.

First of all, on the pilot test set, the basic statistical APE method damaged the original MT output quality, with a TER reduction of about 1 point. On this year's data it achieves a small improvement (though statistically significant only in terms of BLEU). This suggests that, as hypothesized in Section 7.1.1, the higher repetitiveness featured by the selected data can facilitate the work of the APE systems. The new scenario, with repetition rates for SRC, TGT and PE that are more than twice the values measured last year (see Table 32), makes them able to learn from the training data a larger number of reliable and re-applicable correction patterns. However, the large improvements obtained this year by the top runs can only be reached by moving from the basic statistical MT backbone shared by all last year's participants to new and more reliable APE solutions. Indeed, its distance from the top-ranked systems has increased from 0.6 up to 3.12 TER points. While on one side it is true that the new data made the task easier, on the other side the deployed solutions and the increased results' distance over the basic statistical APE approach indicate a significant step forward.

In terms of TER and BLEU evaluations, there are minor differences (only for the lower ranked systems) between the two rankings. This confirms that both metrics capture similar linguistic phenomena and the use of n-grams does not show particular advantages.

7.4 System/performance analysis

Differently from the pilot round, in which TER results were more concentrated (the difference between the top and the lowest ranked system was about 1.5 points), this year systems' performance is distributed within an interval of about 7.5 points. Indeed, the two rankings of Table 34 can be seen as composed of three blocks: the best system, the systems scoring around the baselines and the lower performing systems. Trying to go beyond rough TER/BLEU measurements and to shed light on such performance differences, in this section we focus on a more fine-grained analysis of systems' behaviour and the corresponding errors.

7.4.1 System behaviour

A first interesting aspect to analyse is systems' behaviour which, compared to last year, reflects the larger variety of approaches explored. *Does this variety result in major differences in the correction strategies/operations?* To answer this question, we first analysed the submitted runs taking into consideration the changes made by each system to the test instances. Table 35 shows the number of modified, improved and deteriorated sentences. It's worth noting that, as observed last year, for all the systems the number of modified sentences 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 7.5

	Modified	Improved	Deteriorated
AMU Primary	1,613	935	374
AMU Contrastive	1475	776	386
FBK Contrastive	640	377	148
FBK Primary	654	384	153
USAAR Primary	421	290	74
USAAR Contrastive	499	314	105
CUNI Primary	498	284	138
(Simard et al., 2007)	700	320	253
DCU Contrastive	407	48	314
JUSAAR Primary	1,521	320	835
JUSAAR Contrastive	1,540	326	837
DCU Primary	797	54	651

Table 35: Number of test sentences modified, improved and deteriorated by each submitted run.

Looking at the numbers in Table 35, it becomes evident that the overall number of modified sentences is considerably larger than in the pilot task. On average, the best run submitted by each team modified 42.5% sentences. This amount is much larger than last year, when the percentage was 18.0%, probably due to the higher repetitiveness of the data which makes possible to learn more reliable and applicable correction rules. The same holds for the average number of improved sentences, which this year is significantly larger (18.7% vs. 11% in the pilot). This trend is confirmed by the performance of our re-implementation of Simard et al. (2007), which modified 35% of the sentences (vs. 26% in the pilot), improving 45% (vs. 11% last year) and deteriorating 36% of them (vs. 61%).

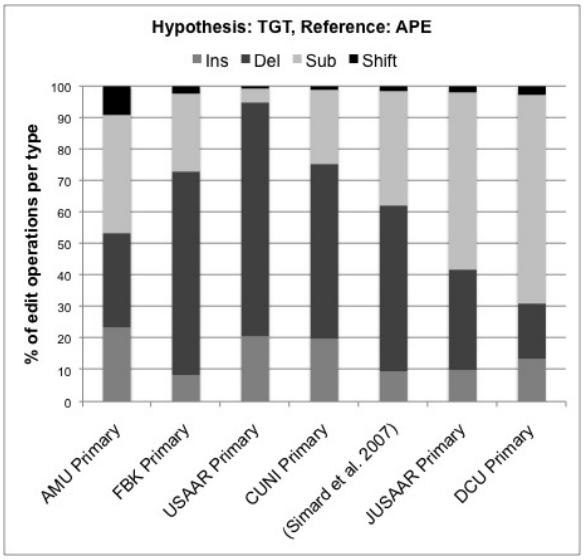

Figure 10: System Behaviour – TER(MT, APE)

These figures, however, vary considerably across the submitted runs. Among the systems that improve over the basic statistical APE approach, the top-ranked one modified an impressive number of test sentences (80%), which is more than twice the amount of items changed by the other submissions. For the same system, the improved and the deteriorated ones are respectively about 58% and 23% of the total, which is in line with the other participants that improved the baseline. An interesting general conclusion that we can draw is that the neural approach adopted by the top-ranked system allowed it to better cope with the data sparsity issues that affect the other methods (despite the higher repetitiveness of this year's data). More thorough investigations that are beyond the scope of this overview should verify the hypothesis that learning and generalising rules from a relatively small amount of human post-edits is easier with neural models than with pure statistical solutions. Another aspect that should be checked is whether the neural solution performs better *per se* or thanks to the much larger amount of training data needed for its deployment.

Further insights about systems' behaviour can be drawn from the analysis of Figure 10. It plots the distribution of the edit operations done by each system (insertions, deletions, substitutions, shifts) obtained by computing the TER between the original MT output and the output of each system as reference (only for the primary submissions).

The figure evidences some interesting trends, starting from the much larger proportion of shifts made by the top-ranked neural approach. More than 450 shift operations (9.2% of the total), in fact, represent the major difference between the behaviour of the winning system and all the

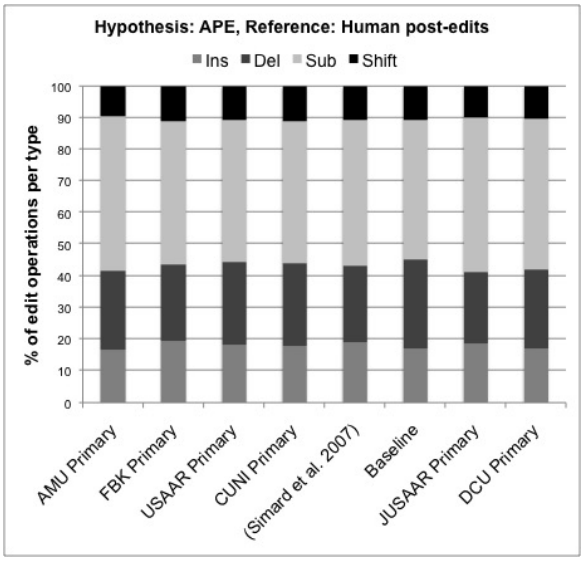

Figure 11: System Error – TER(APE, human post-edits)

other submissions (the second-ranked one performs only 26 shifts, 2.5% of the total). It is likely, but this should be verified, that the available training data featured correction patterns that the neural method was able to model and re-apply better than the other solutions. Overall, the behaviour of the best system is the most balanced with respect the three other operations. In total, insertions, deletions and substitutions (respectively 1,132, 1,465 and 1,807) are considerably more that those made by the other systems and they are more evenly distributed (23%, 30% and 37% respectively). As a term of comparison, the second-ranked primary submission performed much less operations (83 insertions, 652 deletions and 248 substitutions), with a clear predominance (65%) of deletions that is common also to other submissions. As a general remark, best results seem to be associated with a rather homogeneous distribution of the types of correction patterns learned by the system.

7.4.2 System error

Another interesting aspect to analyse is the effect of the different methods on the types of errors made by each system. *Does the variety in the approaches result in major differences in the types of errors made?* To answer this question, Figure 11 plots the distribution of the edit operations needed to transform the output of each system into the human post-edits available for each test sentence. Such distribution of systems' errors is obtained by computing the TER between their output and the human post-edits of the original translations as reference.

The figure does not show visible trends that can provide us with useful hints. In terms of error distribution, the task baseline, our re-implementation of Simard et al. (2007), and the submitted primary runs show almost identical ratios. Insertions range between 17% and 20% of the total, deletions range between 23% and 28%, substitutions range between 44% and 49%. The highest percentage of substitution errors suggests that the major problem for all systems is the lexical choice. Half of the errors in the APE output belong to this error category, indicating that learning the appropriate lexical replacements from human post-edits is still one of the main challenges. Comparing the error distribution in the MT baseline (our ground truth in terms of what has to be corrected) with the actions actually made by each system as shown in Figure 10, it is interesting to emphasise the higher similarity with the distributions of the operations made by the top-performing system. "AMU Primary", indeed, seems to perform a slightly larger amount of insertions compared to the total insertions actually needed, while the other operations are substantially in line with the expected amount. Based on TER information, nothing can be said about which of them are actually correct/wrong. The only conclusions we can draw at this stage are: *i)* a good amount of MT errors is corrected (the global TER decreases), *ii)* the actions of the top-performing system are quite evenly distributed, *iii)* such distribution is the closest to the distribution of ground truth operations but *iv)* errors (missing corrections and/or wrong corrections) still remain in all classes.

In light of these considerations, we performed further analysis by evaluating this years' APE submissions also from another point of view. To this aim, in the next section we try to understand the relation between the participants' performance and the human perception of translation quality.

7.5 Human Evaluation

To assess the quality of APE systems and produce a ranking based on human judgement, as well as analyze how humans perceive TER/BLEU performance differences between the submitted systems, two runs of human evaluations were conducted. The whole evaluation took approximately a month and was performed mainly by student translators who annotated the APE systems' outputs. This subsection describes the human evaluation pro-

cedure, gives details about the annotators' backgrounds and profiles, and finally presents the results of the evaluation.

7.5.1 Evaluation Procedure

The two runs of human evaluation were conducted using the Appraise[37] (Federmann, 2012) open-source annotation platform through the *ranking task* interface. A ranking task consists of a source segment and the outputs of up to 5 anonymized APE systems randomly selected from the set of participants and displayed in random order to human evaluators. The main difference between the two evaluation runs is the following: for the first run, the annotators were presented with a translation reference consisting of the manual post-edit of the machine-translated source segment, while for the second run no translation reference was presented to the human evaluator. For both evaluation runs, the non-post-edited MT output was included among the systems to evaluate. For the second evaluation run, the human post-edited version of the MT output was included among the systems to evaluate.

A total of 200 randomly extracted source segments taken from the test set presented in Table 31 with their corresponding systems' outputs were considered for the first evaluation run, while 100 source segments went through the second run. The decision to consider a larger set of segments for the first evaluation run is based on the previous editions of WMT, where human evaluations conducted for the translation tasks included a translation reference. The smaller scale evaluation for the second run can be seen as a pilot study, where no translation reference is given to the annotators and where the human post-edit is presented as part of the anonymized systems. The latter setup allows us to see if APE systems can reach human post-editing in terms of quality while avoiding evaluation bias towards a reference.

We carried out six annotation sessions in a controlled environment of approximately 45 to 60 minutes each, divided in two blocks of equal duration with a small break in between. Prior to the human evaluation task, we provided annotators with a pilot study in order to be introduced to the ranking task and be familiarized with the annotation interface. For each source sentence, five systems' outputs were randomly selected among the participants and the non-post-edited MT output. For the second evaluation run, the human post-edit was included in the random selection of target sentences to annotate. The human annotators then ranked the outputs from 1 to 5 (1 being the best) with ties allowed. All source segments were evaluated by at least 3 annotators. The annotations were then used with the TrueSkill[38] adaptive ranking system to produce a score for each system based on their inferred means (Sakaguchi et al., 2014). This score was used to sort and cluster the systems submitted by the participants, as well as the MT output and the human post-edit, and produce the final ranking presented in Section 7.5.3

7.5.2 Annotators Background

A total of 37 annotators participated in the manual evaluation of APE systems, including 30 5$^{\text{th}}$ semester B.A. students in the *Comparative Linguistics, Literature, and Translation* program taught in Saarland University.[39] The remaining 7 evaluators are expert translators and lecturers at Saarland University in the *Applied Linguistics, Translation and Interpreting* department.[40] Among the annotators, 34 are native German speakers with strong English skills and have completed introductory courses such as translation theory and translation studies, machine translation, CAT tools, and MT evaluation and post-editing. The remaining 3 annotators have strong German skills and have been living in Germany for several years.

7.5.3 Results

The first and second runs of human evaluation results are respectively presented in Table 36 and Table 37.

The first run shows a preference for the AMU Primary system compared to the other submissions (Table 36). These results confirm those obtained with the automatic metrics as shown in Table 34 and we can see that two systems are above the Baseline (the raw MT output). The CUNI Primary and USAAR Primary systems are in the same cluster with the Baseline, which indicates a non-significant difference with $p \leq 0.05$. Two systems are in a single cluster below the baseline, namely JUSAAR Primary and DCU Primary, being on par with the results obtained using au-

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

[38] https://github.com/keisks/wmt-trueskill

[39] http://fr46.uni-saarland.de/?id=2393

[40] http://fr46.uni-saarland.de

#	Score	Range	ID
1	1.967	1	AMU Primary
2	0.033	2	FBK Primary
3	-0.108	3-4	CUNI Primary
	-0.191	3-5	USAAR Primary
	-0.211	3-5	Baseline
4	-0.712	6-7	JUSAAR Primary
	-0.778	6-7	DCU Primary

Table 36: Results of the first run of human evaluation including human post-edited MT output as translation reference. Scores and ranges are obtained with TrueSkill (Sakaguchi et al., 2014). Lines between systems indicate clusters according to bootstrap resampling at p-level $p \leq 0.05$ based on $1,000$ runs. Systems within a cluster are considered tied.

#	Score	Range	ID
1	2.058	1	Human Post-edit
2	0.867	2	AMU Primary
3	-0.213	3-4	CUNI Primary
	-0.348	3-6	FBK Primary
	-0.374	3-6	USAAR Primary
	-0.499	5-7	Baseline
	-0.675	6-8	JUSAAR Primary
	-0.816	7-8	DCU Primary

Table 37: Results of the second run of human evaluation without translation reference provided to annotators. Scores and ranges are obtained with TrueSkill (Sakaguchi et al., 2014). Lines between systems indicate clusters according to bootstrap resampling at p-level $p \leq 0.05$ based on $1,000$ runs. Systems within a cluster are considered tied.

tomatic metrics. The correlation between automatic metrics and the first manual evaluation run indicates the reliability of popular MT metrics for the evaluation of APE systems. On average, annotators needed 53 seconds to perform one ranking task, while the fastest ranking was performed in 18.3 seconds and the slowest one took more than 4 minutes and 30 seconds (averaged over at least 3 annotators for the same source segment). The agreement between annotators on the first run of evaluation is $k = 0.481$ according to Fleiss' Kappa (Fleiss, 1971).

The results of the second run of manual evaluation (Table 37) show that the human post-editing of MT output is preferred by human annotators when compared to the other systems' outputs, reaching the first position. It indicates that, in spite of the significant improvements over the original MT output, none of the submitted APE systems managed to reach the translation quality achieved by human post-editing. The second position in the ranking is reached by the AMU Primary system, while a single cluster is ranked third and contains all the remaining systems as well as the Baseline. This smaller amount of clusters can be due to the limited scale of the second run of manual evaluation involving 100 source segments only, compared with the 200 segments for the first run. However, this second run shows that the AMU Primary system is again preferred by human evaluators compared to the other systems without necessarily being closer to the human post-edited MT output, which is not included as a translation reference, and thus without biasing human judgements. The agreement between annotators for the second run of evaluation is slightly lower compared to the first run, with a Fleiss' Kappa of $k = 0.466$. For both runs, the inter-annotator agreement is considered moderate. On average, the annotators needed 60 seconds per ranking task, while the fastest ranked outputs was completed in 21.7 seconds and the slowest one in 3 minutes.

7.6 Lessons learned and outlook

The objectives of this pilot APE task were to: *i)* improve and stabilize the evaluation framework in view of future rounds, *ii)* analyze the effect on task feasibility of data coming from a narrow domain, *iii)* analyze the effect of post-edits collected from professional translators, *iv)* analyze how humans perceive TER/BLEU performance differences between different systems, *v)* measure the progress made during one year of research on the APE task.

Concerning the first point, no specific issues emerged this year calling for major changes. The overall format, starting from the baselines and the evaluation metrics adopted, will likely be kept also for the next round.

As regards points *ii)* and *iii)* the positive effect of domain-specific data and professional-quality post-edits is evident. Most likely, these favorable conditions for automatic post-editing will be kept as well, also because they represent a more standard translation scenario compared to the generic news domain.

Regarding point *iv)*, an interesting finding of the manual evaluation is a correlation between human judgements and the results obtained with automatic metrics. This confirms the reliability of popular MT metrics, namely BLEU and TER, for APE systems evaluation. Despite the baseline improvements and the significant overall TER/BLEU gains, the feedback from human evaluators regard-

ing the quality of the APE MT segments is not fully positive yet, showing that there is still room for improvement. One explanation for this is probably related to the domain specificity of the data set used for this year's APE shared task. Many segments contain sets of instructions and commands that are used in user manuals of the IT domain and were given to annotators without context. The annotators also pointed out that they considered difficult to rank very similar segments, as most APE systems do not make substantial modifications of the MT output, which results in similar outputs in terms of quality and leads to challenging comparisons for humans. This aspect is emphasized when no translation reference is given to the annotators. In this case, only the top-ranked system emerges as a source of corrections that are significantly better than the baseline (in spite of the impressive TER and BLEU gains, respectively up to -3.24 and +5.54 points).

In terms of progress over the last year, this was a successful follow-up. More participants, some of which new, resulted in a larger variety in the submitted systems. Those pursuing the phrase-based approach that dominated the pilot round managed to improve over this common backbone in different ways. Other teams introduced interesting novelties, bringing also into the APE framework the popularity of neural approaches. The tangible result is represented by the large improvements over the (last year unbeaten) baseline achieved by most of the systems. Such gains indicate the good potential of APE systems to improve MT output in black-box conditions and motivate further research and developments.

Acknowledgments

This work was supported in parts by the MosesCore, QT21, QTLeap, EXPERT and CRACKER projects funded by the European Commission (7th Framework Programme and H2020).

The APE task organizers would also like to thank Jan Niehues for training the KIT system used to produce the MT output, Text&Form for producing the manual post-edits, and the annotators involved in the manual evaluation.

References

Abdelsalam, A., Bojar, O., and El-Beltagy, S. (2016). Bilingual Embeddings and Word Alignments for Translation Quality Estimation. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

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

Aires, J., Lopes, G., and Gomes, L. (2016). English-Portuguese Biomedical Translation Task Using a Genuine Phrase-Based Statistical Machine Translation Approach. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Al-Rfou, R., Perozzi, B., and Skiena, S. (2013). Polyglot: Distributed Word Representations for Multilingual NLP. In *Proceedings of the 17th Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria.

Allauzen, A., Aufrant, L., Burlot, F., Lacroix, O., Knyazeva, E., Lavergne, T., Wisniewski, G., and Yvon, F. (2016). LIMSI@WMT16: Machine Translation of News. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Avramidis, E. (2016). DFKI's system for WMT16 IT-domain task, including analysis of systematic errors. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Aziz, W., de Sousa, S. C. M., and Specia, L. (2012). Pet: a tool for post-editing and assessing machine translation. In *Eighth International Conference on Language Resources and Evaluation*, LREC, pages 3982–3987, Instanbul, Turkey.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Béchara, H., Ma, Y., and van Genabith, J. (2011). Statistical Post-Editing for a Statistical MT System. In *Proceedings of the 13th Machine Translation Summit*, pages 308–315, Xiamen, China.

Beck, D., Vlachos, A., Paetzold, G., and Specia, L. (2016). SHEF-MIME: Word-level Quality

Estimation Using Imitation Learning. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Bektaş, E., Yilmaz, E., Mermer, C., and Durgar El-Kahlout, . (2016). TÜBTAK SMT System Submission for WMT2016. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Berard, A., Servan, C., Pietquin, O., and Besacier, L. (2016). MultiVec: a Multilingual and Multilevel Representation Learning Toolkit for NLP. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Biçici, E. and Way, A. (2015). Referential translation machines for predicting semantic similarity. *Language Resources and Evaluation*, pages 1–27.

Bicici, E. (2016a). ParFDA for Instance Selection for Statistical Machine Translation. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Bicici, E. (2016b). Referential Translation Machines for Predicting Translation Performance. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Blain, F., Song, X., and Specia, L. (2016). Sheffield Systems for the English-Romanian WMT Translation Task. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (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.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. (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.

Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., and Turchi, M. (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.

Bojar, O., Dušek, O., Kocmi, T., Libovický, J., Novák, M., Popel, M., Sudarikov, R., and Variš, D. (2016a). 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.

Bojar, O., Ercegovčević, M., Popel, M., and Zaidan, O. (2011). A grain of salt for the wmt manual evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 1–11, Edinburgh, Scotland. Association for Computational Linguistics.

Bojar, O., Graham, Y., , and Stanojević, A. K. M. (2016b). Results of the WMT16 Metrics Shared Task . In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Bradbury, J. and Socher, R. (2016). MetaMind Neural Machine Translation System for WMT 2016. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Buck, C., Heafield, K., and Van Ooyen, B. (2014). N-gram counts and language models from the common crawl. *LREC*, 2:4.

Buck, C. and Koehn, P. (2016). Findings of the WMT 2016 Bilingual Document Alignment Shared Task. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

C. de Souza, J. G., Buck, C., Turchi, M., and Negri, M. (2013). FBK-UEdin participation to the WMT13 Quality Estimation shared-task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 352–358.

C. de Souza, J. G., González-Rubio, J., Buck, C., Turchi, M., and Negri, M. (2014). FBK-UPV-UEdin participation in the WMT14 Quality Estimation shared-task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, Baltimore, MD, USA.

Callison-Burch, C., Fordyce, C. S., Koehn, P., Monz, C., and Schroeder, J. (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.

Callison-Burch, C., Fordyce, C. S., Koehn, P., Monz, C., and Schroeder, J. (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.

Callison-Burch, C., Koehn, P., Monz, C., Peterson, K., Przybocki, M., and Zaidan, O. (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.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (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.

Callison-Burch, C., Koehn, P., Monz, C., and Schroeder, J. (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.

Callison-Burch, C., Koehn, P., Monz, C., and Zaidan, O. (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.

Chatterjee, R., C. de Souza, J. G., Negri, M., and Turchi, M. (2016). The FBK Participation in the WMT 2016 Automatic Post-editing Shared Task. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

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

Chatterjee, R., Weller, M., Negri, M., and Turchi, M. (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.

Chung, J., Cho, K., and Bengio, Y. (2016). NYU-MILA Neural Machine Translation Systems for WMT16. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Costa-jussà, M. R., España Bonet, C., Madhyastha, P., Escolano, C., and Fonollosa, J. A. R. (2016). The TALP–UPC Spanish–English WMT Biomedical Task: Bilingual Embeddings and Char-based Neural Language Model Rescoring in a Phrase-based System. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Cuong, H., Frank, S., and Sima'an, K. (2016). ILLC-UvA Adaptation System (Scorpio) at WMT'16 IT-DOMAIN Task. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Ding, S., Duh, K., Khayrallah, H., Koehn, P., and Post, M. (2016). The JHU Machine Translation Systems for WMT 2016. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Duma, M.-S. and Menzel, W. (2016). Data Selection for IT Texts using Paragraph Vector. In *Proceedings of the First Conference on Machine*

Translation, Berlin, Germany. Association for Computational Linguistics.

Durrani, N., Schmid, H., and Fraser, A. (2011). A joint sequence translation model with integrated reordering. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1045–1054. Association for Computational Linguistics.

Dušek, O., Gomes, L., Novák, M., Popel, M., and Rosa, R. (2015). New Language Pairs in TectoMT. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 98–104, Lisbon, Portugal. Association for Computational Linguistics.

Dvorkovich, A., Gubanov, S., and Galinskaya, I. (2016). Yandex School of Data Analysis approach to English-Turkish translation at WMT16 News Translation Task. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Esplà-Gomis, M., Sánchez-Martínez, F., and Forcada, M. (2015). UAlacant word-level machine translation quality estimation system at WMT 2015. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 309–315, Lisbon, Portugal.

Esplà-Gomis, M., Sánchez-Martínez, F., and Forcada, M. (2016). UAlacant word-level and phrase-level machine translation quality estimation systems at WMT 2016. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

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

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Gao, Q. and Vogel, S. (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.

Gaudio, R., Labaka, G., Agirre, E., Osenova, P., Simov, K., Popel, M., Oele, D., van Noord, G., Gomes, L., António Rodrigues, J. a., Neale, S., Silva, J. a., Querido, A., Rendeiro, N., and Branco, A. (2016). SMT and Hybrid systems of the QTLeap project in the WMT16 IT-task. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Giménez, J. and Màrquez, L. (2010). Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, (94):77–86.

Graham, Y. (2015). Improving Evaluation of Machine Translation Quality Estimation. 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*, pages 1804–1813, Beijing, China.

Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2013). Continuous Measurement Scales in Human Evaluation of Machine Translation. In *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (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.

Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2016). Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, pages 1–28.

Grönroos, S.-A., Virpioja, S., and Kurimo, M. (2016). Hybrid Morphological Segmentation for Phrase-Based Machine Translation. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Guillou, L., Hardmeier, C., Nakov, P., Stymne, S., Tiedemann, J., Versley, Y., Cettolo, M., Webber, B., and Popescu-Belis, A. (2016). Findings of the 2016 WMT Shared Task on Cross-lingual Pronoun Prediction. In *Proceedings of the First Conference on Machine Transla-*

tion, Berlin, Germany. Association for Computational Linguistics.

Gwinnup, J., Anderson, T., Erdmann, G., Young, K., Kazi, M., Salesky, E., and Thompson, B. (2016). The AFRL-MITLL WMT16 News-Translation Task Systems. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Ha, T.-L., Cho, E., Niehues, J., Mediani, M., Sperber, M., Allauzen, A., and Waibel, A. (2016). The Karlsruhe Institute of Technology Systems for the News Translation Task in WMT 2016. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Heafield, K. (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.

Herrmann, T., Niehues, J., and Waibel, A. (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.

Hoang, C. and Sima'an, K. (2014). Latent domain translation models in mix-of-domains haystack. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1928–1939, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Huck, M., Fraser, A., and Haddow, B. (2016). The Edinburgh/LMU Hierarchical Machine Translation System for WMT 2016. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Ive, J., Max, A., and Yvon, F. (2016). LIMSI's Contribution to the WMT'16 Biomedical Translation Task. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Jawaid, B., Kamran, A., Stanojević, M., and ojar, O. (2016). Results of the WMT16 Tuning Shared Task. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Junczys-Dowmunt, M., Dwojak, T., and Sennrich, R. (2016). The AMU-UEDIN Submission to the WMT16 News Translation Task: Attention-based NMT Models as Feature Functions in Phrase-based SMT. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Junczys-Dowmunt, M. and Grundkiewicz, R. (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*. Association for Computational Linguistics.

Kim, H. and Lee, J.-H. (2016). Recurrent Neural Network based Translation Quality Estimation. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Koehn, P. (2002). Europarl: A multilingual corpus for evaluation of machine translation. Unpublished, `http://www.isi.edu/~koehn/europarl/`.

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

Koehn, P. and Hoang, H. (2007). Factored Translation Models. In *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 868–876, Prague, Czech Republic.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (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.

Koehn, P. and Monz, C. (2006). Manual and automatic evaluation of machine translation be-

tween european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City. Association for Computational Linguistics.

Kozlova, A., Shmatova, M., and Frolov, A. (2016). YSDA Participation in the WMT'16 Quality Estimation Shared Task. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Kreutzer, J., Schamoni, S., and Riezler, S. (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 297–303, Lisboa, Portugal. Association for Computational Linguistics.

Landis, J. R. and Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33:159–174.

Liang, P., Taskar, B., and Klein, D. (2006). Alignment by agreement. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 104–111. Association for Computational Linguistics.

Libovický, J., Helcl, J., Tlustý, M., Bojar, O., and Pecina, P. (2016). CUNI at Post-editing and Multimodal Translation Tasks. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics.

Lo, C.-k., Cherry, C., Foster, G., Stewart, D., Islam, R., Kazantseva, A., and Kuhn, R. (2016). NRC Russian-English Machine Translation System for WMT 2016. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Logacheva, V., , and Specia, L. (2015). Phrase-level Quality Estimation for Machine Translation. In *Proceedings of the 12th International Workshop on Spoken Language Translation*, Da Nang, Vietnam.

Logacheva, V., Blain, F., and Specia, L. (2016a). USFD Phrase-level Quality Estimation Systems. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Logacheva, V., Hokamp, C., and Specia, L. (2016b). 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.

Logacheva, V., Lukasik, M., and Specia, L. (2016c). Metrics for Evaluation of Word-Level Machine Translation Quality Estimation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany.

Luong, N. Q., Besacier, L., and Lecouteux, B. (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.

Mareček, D. (2016). Merged bilingual trees based on Universal Dependencies in Machine Translation. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Martins, A., Almeida, M., and Smith, N. (2013). Turning on the turbo: Fast third-order non-projective turbo parsers. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 617–622, Sofia, Bulgaria.

Martins, A. F. T., Astudillo, R., Hokamp, C., and Kepler, F. (2016). Unbabel's Participation in the WMT16 Word-Level Translation Quality Estimation Shared Task. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Mikolov, T., Yih, W.-t., and Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *Proceedings of NAACL 2013*.

Molchanov, A. and Bykov, F. (2016). PROMT Translation Systems for WMT 2016 Translation Tasks. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Neves, M., Yepes, A. J., and Névéol, A. (2016). The Scielo Corpus: a Parallel Corpus of Scientific Publications for Biomedicine. In *Proceedings of the Tenth International Conference*

on Language Resources and Evaluation (LREC 2016), Paris, France. European Language Resources Association (ELRA).

Och, F. J. (2003). Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Sapporo, Japan.

Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Comput. Linguist.*, 29(1):19–51.

Okazaki, N. (2007). CRFsuite: a fast implementation of Conditional Random Fields.

Paetzold, G. and Specia, L. (2016). SimpleNets: Quality Estimation with Resource-Light Neural Networks. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Pahari, K., Kuila, A., Pal, S., Naskar, S. K., Bandyopadhyay, S., and van Genabith, J. (2016). JU-USAAR: A Domain Adaptive MT System. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Pal, S., Mihaela, V., Naskar, S. K., and van Genabith, J. (2015). USAAR-SAPE: An English–Spanish Statistical Automatic Post-Editing System. In *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT)*, pages 216–221.

Pal, S., Zampieri, M., and van Genabith, J. (2016). USAAR: An Operation Sequential Model for Automatic Statistical Post-Editing. In *Proceedings of the First Conference on Machine Translation*. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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.

Patel, R. N. and M, S. (2016). Translation Quality Estimation using Recurrent Neural Network. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Perez-de Viñaspre, O. and Labaka, G. (2016). IXA Biomedical Translation System at WMT16 Biomedical Translation Task. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Peter, J.-T., Alkhouli, T., Guta, A., and Ney, H. (2016a). The RWTH Aachen University English-Romanian Machine Translation System for WMT 2016. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Peter, J.-T., Alkhouli, T., Ney, H., Huck, M., Braune, F., Fraser, A., Tamchyna, A., Bojar, O., Haddow, B., Sennrich, R., Blain, F., Specia, L., Niehues, J., Waibel, A., Allauzen, A., Aufrant, L., Burlot, F., knyazeva, e., Lavergne, T., Yvon, F., Daiber, J., and Pinnis, M. (2016b). The QT21/HimL Combined Machine Translation System. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, Massachusetts.

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

Rosa, R., Sudarikov, R., Novák, M., Popel, M., and Bojar, O. (2016). Dictionary-based Domain Adaptation of MT Systems without Retraining. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Sagemo, O. and Stymne, S. (2016). The UU Submission to the Machine Translation Quality Estimation Task. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Sakaguchi, K., Post, M., and Van Durme, B. (2014). Efficient elicitation of annotations for

human evaluation of machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 1–11, Baltimore, Maryland, USA. Association for Computational Linguistics.

Sánchez-Cartagena, V. M. and Toral, A. (2016). Abu-MaTran at WMT 2016 Translation Task: Deep Learning, Morphological Segmentation and Tuning on Character Sequences. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Scarton, C., Beck, D., Shah, K., Sim Smith, K., and Specia, L. (2016). Word embeddings and discourse information for Quality Estimation. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Scarton, C. and Specia, L. (2014). Document-level translation quality estimation: exploring discourse and pseudo-references. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pages 101–108, Dubrovnik, Croatia.

Scarton, C., Tan, L., and Specia, L. (2015a). USHEF and USAAR-USHEF participation in the WMT15 QE shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 317–322, Lisboa, Portugal. Association for Computational Linguistics.

Scarton, C., Zampieri, M., Vela, M., van Genabith, J., and Specia, L. (2015b). Searching for Context: a Study on Document-Level Labels for Translation Quality Estimation. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 121–128, Antalya, Turkey.

Seddah, D., Kübler, S., and Tsarfaty, R. (2014). Introducing the SPMRL 2014 Shared Task on Parsing Morphologically-rich Languages. In *Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages*, pages 103–109, Dublin, Ireland.

Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Sennrich, R., Haddow, B., and Birch, A. (2016). Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Shah, K., Bougares, F., Barrault, L., and Specia, L. (2016). SHEF-LIUM-NN: Sentence level Quality Estimation with Neural Network Features. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Sim Smith, K., Aziz, W., and Specia, L. (2016). Cohere: A Toolkit for Local Coherence. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, Portorož, Slovenia.

Simard, M., Goutte, C., and Isabelle, P. (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.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, USA.

Specia, L., Frank, S., Sima'an, K., and Elliott, D. (2016). A Shared Task on Multimodal Machine Translation and Crosslingual Image Description. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Specia, L., Paetzold, G., and Scarton, C. (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.

Stahlberg, F., Hasler, E., and Byrne, B. (2016). The Edit Distance Transducer in Action: The University of Cambridge English-German System at WMT16. In *Proceedings of the First Conference on Machine Translation*, Berlin,

Germany. Association for Computational Linguistics.

Sudarikov, R., Popel, M., Bojar, O., Burchardt, A., and Klejch, O. (2016). Using MT-ComparEval. In *Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem*, pages 76–82.

Tamchyna, A., Sudarikov, R., Bojar, O., and Fraser, A. (2016). CUNI-LMU Submissions in WMT2016: Chimera Constrained and Beaten. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Tezcan, A., Hoste, V., and Macken, L. (2016). UGENT-LT3 SCATE Submission for WMT16 Shared Task on Quality Estimation. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

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

Tiedemann, J., Cap, F., Kanerva, J., Ginter, F., Stymne, S., Östling, R., and Weller-Di Marco, M. (2016). Phrase-Based SMT for Finnish with More Data, Better Models and Alternative Alignment and Translation Tools. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Williams, P., Sennrich, R., Nadejde, M., Huck, M., Haddow, B., and Bojar, O. (2016). Edinburgh's Statistical Machine Translation Systems for WMT16. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Wołk, K. and Marasek, K. (2016). PJAIT Systems for the WMT 2016. In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Yeh, A. (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.

A Pairwise System Comparisons by Human Judges

Tables 40–46 show pairwise comparisons between systems for each language pair. The numbers in each of the tables' cells indicate the percentage of times that the system in that column was judged to be better than the system in that row, ignoring ties. Bolding indicates the winner of the two systems.

Because there were so many systems and data conditions the significance of each pairwise comparison needs to be quantified. We applied the Sign Test to measure which comparisons indicate genuine differences (rather than differences that are attributable to chance). In the following tables ⋆ indicates statistical significance at $p \leq 0.10$, † indicates statistical significance at $p \leq 0.05$, and ‡ indicates statistical significance at $p \leq 0.01$, according to the Sign Test.

Each table contains final rows showing how likely a system would win when paired against a randomly selected system (the expected win ratio score) and the rank range according bootstrap resampling ($p \leq 0.05$). Gray lines separate clusters based on non-overlapping rank ranges.

	ONLINE-B	UEDIN-NMT	UEDIN-PBMT	UEDIN-SYNTAX	ONLINE-A	JHU-PBMT	LIMSI
ONLINE-B	–	.47⋆	.43‡	.39‡	.39‡	.38‡	.36‡
UEDIN-NMT	**.53**⋆	–	.45‡	.43‡	.41‡	.40‡	.39‡
UEDIN-PBMT	**.57**‡	**.55**‡	–	.46†	.45‡	.39‡	.41‡
UEDIN-SYNTAX	**.61**‡	**.57**‡	**.54**†	–	.49	.44‡	.44‡
ONLINE-A	**.61**‡	**.59**‡	**.55**‡	**.51**	–	.47⋆	.47⋆
JHU-PBMT	**.62**‡	**.60**‡	**.61**‡	**.56**‡	**.53**⋆	–	.46†
LIMSI	**.64**‡	**.61**‡	**.59**‡	**.56**‡	**.53**⋆	**.54**†	–
score	.58	.37	.09	-.08	-.18	-.32	-.46
rank	1-2	1-2	3	4-5	4-6	5-7	6-7

Table 38: Head to head comparison, ignoring ties, for Romanian-English systems

	UEDIN-NMT	QT21-HIML-SYSCOMB	KIT	UEDIN-PBMT	ONLINE-B	UEDIN-LMU-HIERO	RWTH-SYSCOMB	LIMSI	LMU-CUNI	JHU-PBMT	USFD-RESCORING	ONLINE-A
UEDIN-NMT	–	.48	.43⋆	.40†	.36‡	.42†	.38‡	.31‡	.37‡	.34‡	.28‡	.25‡
QT21-HIML-SYSCOMB	**.52**	–	.44	.41†	.44	.40†	.41†	.30‡	.25‡	.28‡	.22‡	.22‡
KIT	**.57**⋆	**.56**	–	**.52**	.44	.47	.43⋆	.36‡	.35‡	.41†	.33‡	.34‡
UEDIN-PBMT	**.60**†	**.59**†	.48	–	.49	.47	**.57**⋆	.39‡	.36‡	.32‡	.32‡	.34‡
ONLINE-B	**.64**‡	**.56**	**.56**	**.51**	–	.49	.49	.41†	.37‡	.35‡	.28‡	.36‡
UEDIN-LMU-HIERO	**.58**†	**.60**†	**.53**	**.53**	**.51**	–	.50	.43⋆	.37‡	.38‡	.30‡	.29‡
RWTH-SYSCOMB	**.62**‡	**.59**†	**.57**⋆	.43⋆	**.51**	.50	–	.42⋆	.38‡	.42⋆	.34‡	.31‡
LIMSI	**.69**‡	**.70**‡	**.64**‡	**.61**‡	**.59**†	**.57**⋆	**.58**⋆	–	.48	.43⋆	.47	.35‡
LMU-CUNI	**.63**‡	**.75**‡	**.65**‡	**.64**‡	**.63**‡	**.63**‡	**.62**‡	**.52**	–	**.52**	.42†	.40†
JHU-PBMT	**.66**‡	**.72**‡	**.59**†	**.68**‡	**.65**‡	**.62**‡	**.58**⋆	**.57**⋆	.48	–	.50	.42†
USFD-RESCORING	**.72**‡	**.78**‡	**.67**‡	**.68**‡	**.72**‡	**.70**‡	**.66**‡	**.53**	**.58**†	.50	–	.39‡
ONLINE-A	**.75**‡	**.78**‡	**.66**‡	**.66**‡	**.64**‡	**.71**‡	**.69**‡	**.65**‡	**.60**†	**.58**†	**.61**‡	–
score	.44	.43	.20	.15	.14	.13	.12	-.15	-.22	-.26	-.43	-.56
rank	1-2	1-2	3-7	3-7	3-7	3-7	3-7	8-10	8-10	8-11	10-12	11-12

Table 39: Head to head comparison, ignoring ties, for English-Romanian systems

	UEDIN-NMT	JHU-PBMT	ONLINE-B	TT-BLEU-MIRA	TT-AFRL	TT-NRC-NNBLEU	TT-NRC-MEANT	TT-BEER-PRO	PJATK	TT-BLEU-MERT	ONLINE-A	CU-MERGEDTREES
UEDIN-NMT	–	.42‡	.41‡	.36‡	.36‡	.37‡	.35‡	.35‡	.35‡	.36‡	.33‡	.14‡
JHU-PBMT	**.58**‡	–	.45‡	.45‡	.43‡	.44‡	.42‡	.40‡	.41‡	.40‡	.38‡	.13‡
ONLINE-B	**.59**‡	**.55**‡	–	.47⋆	.46†	.46‡	.45‡	.45‡	.44‡	.42‡	.43‡	.16‡
TT-BLEU-MIRA	**.64**‡	**.55**‡	**.53**⋆	–	.49	.47†	.47†	.45‡	.45‡	.42‡	.45‡	.15‡
TT-AFRL	**.64**‡	**.57**‡	**.54**†	**.51**	–	.49	.47†	.43‡	.46‡	.45‡	.44‡	.16‡
TT-NRC-NNBLEU	**.63**‡	**.56**‡	**.54**‡	**.53**†	**.51**	–	.50	.46‡	.47‡	.43‡	.46†	.16‡
TT-NRC-MEANT	**.65**‡	**.58**‡	**.55**‡	**.53**†	**.53**†	.50	–	.46†	.48†	.47‡	.45‡	.15‡
TT-BEER-PRO	**.65**‡	**.60**‡	**.55**‡	**.55**‡	**.57**‡	**.54**‡	**.54**†	–	.49	.49	.47⋆	.17‡
PJATK	**.65**‡	**.59**‡	**.56**‡	**.55**‡	**.54**‡	**.53**‡	**.52**†	**.51**	–	.50	.47⋆	.18‡
TT-BLEU-MERT	**.64**‡	**.60**‡	**.58**‡	**.58**‡	**.55**‡	**.57**‡	**.53**‡	**.51**	.50	–	.48	.19‡
ONLINE-A	**.67**‡	**.62**‡	**.57**‡	**.55**‡	**.56**‡	**.54**†	**.55**‡	**.53**⋆	**.53**⋆	**.52**	–	.19‡
CU-MERGEDTREES	**.86**‡	**.87**‡	**.84**‡	**.85**‡	**.84**‡	**.84**‡	**.85**‡	**.83**‡	**.82**‡	**.81**‡	**.81**‡	–
score	.61	.31	.20	.11	.09	.09	.07	.03	.00	.00	-.07	-.148
rank	1	2	3	4-6	4-7	4-7	5-8	7-10	8-10	8-10	11	12

Table 40: Head to head comparison, ignoring ties, for Czech-English systems

	UEDIN-NMT	NYU-UMONTREAL	JHU-PBMT	CU-CHIMERA	CU-TAMCHYNA	UEDIN-CU-SYNTAX	ONLINE-B	TT-BLEU-MIRA	TT-BEER-PRO	TT-BLEU-MERT	TT-AFRL2	TT-AFRL1	TT-DCU	TT-FJFI	ONLINE-A	CU-TECTOMT	TT-USAAR-HMM-MERT	CU-MERGEDTREES	TT-USAAR-HMM-MIRA	TT-USAAR-HARMONIC
UEDIN-NMT	–	.38‡	.31‡	.33‡	.33‡	.35‡	.31‡	.26‡	.25‡	.27‡	.22‡	.25‡	.28‡	.26‡	.21‡	.20‡	.11‡	.07‡	.00‡	.01‡
NYU-MONTREAL	**.62**‡	–	.43‡	.42‡	.41‡	.37‡	.33‡	.38‡	.36‡	.37‡	.34‡	.36‡	.31‡	.37‡	.30‡	.21‡	.14‡	.09‡	.01‡	.00‡
JHU-PBMT	**.69**‡	**.57**‡	–	.45‡	.47†	.47	.38‡	.37‡	.37‡	.38‡	.36‡	.35‡	.35‡	.36‡	.35‡	.28‡	.10‡	.12‡	.01‡	.00‡
CU-CHIMERA	**.67**‡	**.58**‡	**.55**‡	–	.49	.46⋆	.43‡	.40‡	.39‡	.40‡	.39‡	.39‡	.40‡	.39‡	.39‡	.30‡	.12‡	.10‡	.01‡	.00‡
CU-TAMCHYNA	**.67**‡	**.59**‡	**.53**†	**.51**	–	.45†	.42‡	.41‡	.41‡	.40‡	.40‡	.40‡	.39‡	.38‡	.39‡	.29‡	.16‡	.11‡	.01‡	.00‡
UEDIN-CU-SNTX	**.65**‡	**.63**‡	**.53**	**.54**⋆	**.54**†	–	.49	.48	.47	.47⋆	.49	.45†	.46†	.44‡	.40‡	.37‡	.16‡	.14‡	.01‡	.00‡
ONLINE-B	**.69**‡	**.67**‡	**.62**‡	**.57**‡	**.58**‡	**.51**	–	.48⋆	.46‡	.48†	.44‡	.44‡	.48⋆	.46‡	.41‡	.38‡	.15‡	.12‡	.01‡	.00‡
TT-BLEU-MIRA	**.74**‡	**.62**‡	**.63**‡	**.60**‡	**.59**‡	**.52**	**.52**⋆	–	.49	.46⋆	.46†	.46†	.43‡	.47⋆	.43‡	.39‡	.12‡	.13‡	.01‡	.00‡
TT-BEER-PRO	**.75**‡	**.64**‡	**.63**‡	**.61**‡	**.59**‡	**.53**	**.54**‡	**.51**	–	**.51**	.47	.47⋆	.46†	.47†	.46⋆	.40‡	.14‡	.11‡	.01‡	.00‡
TT-BLEU-MERT	**.73**‡	**.63**‡	**.62**‡	**.60**‡	**.60**‡	**.53**⋆	**.52**†	**.54**⋆	.49	–	.48	.48	.48	.48	.44‡	.39‡	.11‡	.14‡	.01‡	.00‡
TT-AFRL2	**.78**‡	**.66**‡	**.64**‡	**.61**‡	**.60**‡	**.51**	**.56**‡	**.54**†	**.53**	**.52**	–	.47	.48⋆	.48	.43‡	.42‡	.14‡	.11‡	.00‡	.00‡
TT-AFRL1	**.75**‡	**.64**‡	**.65**‡	**.61**‡	**.61**‡	**.55**†	**.56**‡	**.54**†	**.53**⋆	**.52**	**.53**	–	.48	.49	.45†	.42‡	.14‡	.10‡	.00‡	.00‡
TT-DCU	**.72**‡	**.69**‡	**.65**‡	**.60**‡	**.61**‡	**.54**†	**.52**⋆	**.57**‡	**.54**†	**.52**	**.52**⋆	**.52**	–	**.51**	.42‡	.44‡	.12‡	.14‡	.01‡	.00‡
TT-FJFI	**.74**‡	**.63**‡	**.64**‡	**.61**‡	**.62**‡	**.56**‡	**.54**‡	**.53**⋆	**.53**†	**.52**	**.52**	**.51**	.49	–	.47	.44‡	.13‡	.15‡	.01‡	.00‡
ONLINE-A	**.79**‡	**.70**‡	**.65**‡	**.61**‡	**.61**‡	**.60**‡	**.59**‡	**.57**‡	**.54**⋆	**.56**‡	**.57**‡	**.55**†	**.58**‡	**.53**	–	.42‡	.20‡	.15‡	.03‡	.00‡
CU-TECTOMT	**.80**‡	**.79**‡	**.72**‡	**.70**‡	**.71**‡	**.63**‡	**.62**‡	**.61**‡	**.60**‡	**.61**‡	**.58**‡	**.58**‡	**.56**‡	**.56**‡	**.58**‡	–	.29‡	.23‡	.02‡	.00‡
TT-US'R-'-MERT	**.89**‡	**.86**‡	**.90**‡	**.88**‡	**.84**‡	**.84**‡	**.85**‡	**.88**‡	**.86**‡	**.89**‡	**.86**‡	**.86**‡	**.88**‡	**.87**‡	**.80**‡	**.71**‡	–	.49	.05‡	.01‡
CU-MTREES	**.93**‡	**.91**‡	**.88**‡	**.90**‡	**.89**‡	**.86**‡	**.88**‡	**.87**‡	**.89**‡	**.86**‡	**.89**‡	**.90**‡	**.86**‡	**.85**‡	**.85**‡	**.77**‡	**.51**	–	.04‡	.00‡
TT-US'R-MIRA	**.100**‡	**.99**‡	**.99**‡	**.99**‡	**.99**‡	**.99**‡	**.99**‡	**.99**‡	**.99**‡	**.99**‡	**.100**‡	**.100**‡	**.99**‡	**.99**‡	**.97**‡	**.98**‡	**.95**‡	**.96**‡	–	.07‡
TT-US'R-HARM	**.99**‡	**.100**‡	**.100**‡	**.100**‡	**.100**‡	**.100**‡	**.100**‡	**.100**‡	**.100**‡	**.100**‡	**.100**‡	**.100**‡	**.100**‡	**.100**‡	**.100**‡	**.100**‡	**.99**‡	**.100**‡	**.93**‡	–
score	.59	.42	.34	.30	.30	.22	.19	.16	.15	.15	.13	.13	.13	.12	.07	-.02	-.43	-.54	-.113	-.132
rank	1	2	3	4-5	4-5	6-7	6-7	8-11	8-12	8-13	9-14	9-14	9-14	11-14	15	16	17	18	19	20

Table 41: Head to head comparison, ignoring ties, for English-Czech systems

	UEDIN-NMT	ONLINE-B	ONLINE-A	UEDIN-SYNTAX	KIT	UEDIN-PBMT	JHU-PBMT	ONLINE-G	JHU-SYNTAX	ONLINE-F
UEDIN-NMT	–	.38‡	.34‡	.36‡	.34‡	.34‡	.32‡	.31‡	.19‡	.21‡
ONLINE-B	**.62**‡	–	.50	.48	.49	.44†	.43‡	.40‡	.30‡	.28‡
ONLINE-A	**.66**‡	.50	–	**.52**	.48	.44†	.44†	.44†	.32‡	.25‡
UEDIN-SYNTAX	**.64**‡	**.52**	.48	–	.50	.46⋆	.47	.40‡	.29‡	.29‡
KIT	**.66**‡	**.51**	**.52**	.50	–	.45†	.47	.43‡	.31‡	.27‡
UEDIN-PBMT	**.66**‡	**.56**†	**.56**†	**.54**⋆	**.55**†	–	.48	.44‡	.33‡	.31‡
JHU-PBMT	**.68**‡	**.57**‡	**.56**†	**.53**	**.53**	**.52**	–	.47	.31‡	.29‡
ONLINE-G	**.69**‡	**.60**‡	**.56**†	**.60**‡	**.57**‡	**.56**‡	**.53**	–	.37‡	.34‡
JHU-SYNTAX	**.81**‡	**.70**‡	**.68**‡	**.71**‡	**.69**‡	**.67**‡	**.69**‡	**.63**‡	–	.50
ONLINE-F	**.79**‡	**.72**‡	**.75**‡	**.71**‡	**.73**‡	**.69**‡	**.71**‡	**.66**‡	.50	–
score	.81	.25	.21	.19	.17	.04	.02	-.12	-.67	-.93
rank	1	2-5	2-5	2-5	2-6	5-7	6-7	8	9	10

Table 42: Head to head comparison, ignoring ties, for German-English systems

	UEDIN-NMT	METAMIND	UEDIN-SYNTAX	NYU-UMONTREAL	ONLINE-B	KIT-LIMSI	CAMBRIDGE	ONLINE-A	PROMT-RULE-BASED	KIT	JHU-SYNTAX	JHU-PBMT	UEDIN-PBMT	ONLINE-F	ONLINE-G
UEDIN-NMT	–	.46	.34‡	.41‡	.31‡	.31‡	.31‡	.29‡	.32‡	.27‡	.27‡	.31‡	.28‡	.25‡	.22‡
METAMIND	**.54**	–	.41‡	.40‡	.33‡	.36‡	.35‡	.35‡	.34‡	.33‡	.29‡	.34‡	.30‡	.29‡	.30‡
UEDIN-SYNTAX	**.66**‡	**.59**‡	–	.44†	.35‡	.39‡	.35‡	.33‡	.41‡	.38‡	.27‡	.36‡	.25‡	.27‡	.27‡
NYU-UMONTREAL	**.59**‡	**.60**‡	**.56**†	–	.39‡	.48	.41‡	.45⋆	.41‡	.44†	.37‡	.39‡	.38‡	.35‡	.34‡
ONLINE-B	**.69**‡	**.67**‡	**.65**‡	**.61**‡	–	.49	**.51**	.49	.49	.48	.46†	.42‡	.38‡	.38‡	.32‡
KIT-LIMSI	**.69**‡	**.64**‡	**.61**‡	**.52**	**.51**	–	**.53**	.48	.50	.45	.47	.42‡	.39‡	.42‡	.43†
CAMBRIDGE	**.69**‡	**.65**‡	**.65**‡	**.59**‡	.49	.47	–	.47	**.53**⋆	.46⋆	.42‡	.48	.39‡	.43‡	.42‡
ONLINE-A	**.71**‡	**.65**‡	**.67**‡	**.55**⋆	**.51**	**.52**	**.53**	–	.47	.49	.47⋆	.44‡	.38‡	.37‡	.36‡
PROMT-RULE-BASED	**.68**‡	**.66**‡	**.59**‡	**.59**‡	**.51**	.50	.47⋆	**.53**	–	.48	.46†	.47⋆	.42‡	.39‡	.41‡
KIT	**.73**‡	**.67**‡	**.62**‡	**.56**†	**.52**	**.55**	**.54**⋆	**.51**	**.52**	–	.46†	.44‡	.40‡	.42‡	.41‡
JHU-SYNTAX	**.73**‡	**.71**‡	**.73**‡	**.63**‡	**.54**†	**.53**	**.58**‡	**.53**⋆	**.54**†	**.54**†	–	.48	.42‡	.46⋆	.42‡
JHU-PBMT	**.69**‡	**.66**‡	**.64**‡	**.61**‡	**.58**‡	**.58**‡	**.52**	**.56**‡	**.53**⋆	**.56**‡	**.52**	–	.43‡	.47	.47
UEDIN-PBMT	**.72**‡	**.70**‡	**.75**‡	**.62**‡	**.62**‡	**.61**‡	**.61**‡	**.62**‡	**.58**‡	**.60**‡	**.58**‡	**.57**‡	–	.45⋆	.48
ONLINE-F	**.75**‡	**.71**‡	**.73**‡	**.65**‡	**.62**‡	**.58**‡	**.57**‡	**.63**‡	**.61**‡	**.58**‡	**.54**⋆	**.53**	**.55**⋆	–	.48
ONLINE-G	**.78**‡	**.70**‡	**.73**‡	**.66**‡	**.68**‡	**.57**†	**.58**‡	**.64**‡	**.59**‡	**.59**‡	**.58**‡	**.53**	**.52**	**.52**	–
score	.49	.39	.28	.16	-.00	-.01	-.02	-.02	-.03	-.04	-.13	-.15	-.25	-.32	-.34
rank	1	2	3	4	5-10	5-10	5-10	5-10	5-10	6-10	11-12	11-12	13-14	13-15	14-15

Table 43: Head to head comparison, ignoring ties, for English-German systems

	UEDIN-PBMT	ONLINE-G	ONLINE-B	UH-OPUS	PROMT-SMT	UH-FACTORED	UEDIN-SYNTAX	ONLINE-A	JHU-PBMT
UEDIN-PBMT	–	.50	.48	.49	.40‡	.36‡	.38‡	.32‡	.21‡
ONLINE-G	.50	–	**.51**	.47⋆	.39‡	.41‡	.38‡	.30‡	.23‡
ONLINE-B	**.52**	.49	–	.50	.39‡	.36‡	.34‡	.35‡	.22‡
UH-OPUS	**.51**	**.53**⋆	.50	–	.42‡	.38‡	.38‡	.34‡	.24‡
PROMT-SMT	**.60**‡	**.61**‡	**.61**‡	**.58**‡	–	.46†	.46†	.42‡	.28‡
UH-FACTORED	**.64**‡	**.59**‡	**.64**‡	**.62**‡	**.54**†	–	.50	.47	.28‡
UEDIN-SYNTAX	**.62**‡	**.62**‡	**.66**‡	**.62**‡	**.54**†	.50	–	.46†	.29‡
ONLINE-A	**.68**‡	**.70**‡	**.65**‡	**.66**‡	**.58**‡	**.53**	**.54**†	–	.34‡
JHU-PBMT	**.79**‡	**.77**‡	**.78**‡	**.76**‡	**.72**‡	**.72**‡	**.71**‡	**.66**‡	–
score	.42	.40	.39	.33	.01	-.11	-.13	-.28	-.102
rank	1-4	1-4	1-4	1-4	5	6-7	6-7	8	9

Table 44: Head to head comparison, ignoring ties, for Finnish-English systems

	ONLINE-G	ABUMATRAN-NMT	ONLINE-B	ABUMATRAN-COMBO	UH-OPUS	ABUMATRAN-PBSMT	NYU-UMONTREAL	ONLINE-A	JHU-PBMT	UH-FACTORED	AALTO	JHU-HLTCOE	UUT
ONLINE-G	–	.50	.49	.47⋆	.46⋆	.38‡	.43‡	.39‡	.33‡	.34‡	.32‡	.30‡	.33‡
ABUMATRAN-NMT	.50	–	.48	.43⋆	.46⋆	.41‡	.43‡	.35‡	.37‡	.38‡	.35‡	.36‡	.34‡
ONLINE-B	**.51**	**.52**	–	.50	.46⋆	.41‡	.40‡	.41‡	.38‡	.35‡	.38‡	.33‡	.31‡
ABUMATRAN-COMBO	**.53⋆**	**.57⋆**	.50	–	.48	.38‡	.45†	.40‡	.38‡	.38‡	.37‡	.37‡	.37‡
UH-OPUS	**.54⋆**	**.54⋆**	**.54⋆**	**.52**	–	.45†	.47	.45†	.42‡	.38‡	.39‡	.39‡	.37‡
ABUMATRAN-PBSMT	**.62‡**	**.59‡**	**.59‡**	**.62‡**	**.55†**	–	.47	**.51**	.47	.42‡	.41‡	.42‡	.41‡
NYU-UMONTREAL	**.57‡**	**.57‡**	**.60‡**	**.55†**	**.53**	**.53**	–	.50	.46⋆	.44‡	.44‡	.45†	.41‡
ONLINE-A	**.61‡**	**.65‡**	**.59‡**	**.60‡**	**.55†**	.49	.50	–	.47	.42‡	.40‡	.37‡	.43‡
JHU-PBMT	**.67‡**	**.63‡**	**.62‡**	**.62‡**	**.58‡**	**.53**	**.54⋆**	**.53**	–	.47	.46⋆	.43‡	.43‡
UH-FACTORED	**.66‡**	**.62‡**	**.65‡**	**.62‡**	**.62‡**	**.58‡**	**.56‡**	**.58‡**	**.53**	–	.49	.46⋆	.47
AALTO	**.68‡**	**.65‡**	**.62‡**	**.63‡**	**.61‡**	**.59‡**	**.56‡**	**.60‡**	**.54⋆**	**.51**	–	**.51**	.46⋆
JHU-HLTCOE	**.70‡**	**.64‡**	**.67‡**	**.63‡**	**.61‡**	**.58‡**	**.55†**	**.62‡**	**.57‡**	**.54⋆**	.49	–	.47⋆
UUT	**.67‡**	**.66‡**	**.69‡**	**.63‡**	**.63‡**	**.59‡**	**.59‡**	**.57‡**	**.57‡**	**.53**	**.54⋆**	**.53⋆**	–
score	.36	.31	.29	.23	.15	-.01	-.01	-.01	-.14	-.22	-.28	-.30	-.35
rank	1-3	1-4	1-4	3-5	4-5	6-8	6-8	6-8	9-10	9-12	10-13	10-13	11-13

Table 45: Head to head comparison, ignoring ties, for English-Finnish systems

	PROMT-RULE-BASED	AMU-UEDIN	ONLINE-B	UEDIN-NMT	ONLINE-G	NYU-UMONTREAL	JHU-PBMT	LIMSI	ONLINE-A	AFRL-MITLL-PHRASE	AFRL-MITLL-VERB-A	ONLINE-F
PROMT-RULE-BASED	–	.38‡	.34‡	.33‡	.33‡	.31‡	.26‡	.31‡	.20‡	.26‡	.21‡	.07‡
AMU-UEDIN	**.62‡**	–	.44†	**.51**	.46⋆	.45⋆	.33‡	.35‡	.32‡	.31‡	.28‡	.14‡
ONLINE-B	**.66‡**	**.56†**	–	.50	.46	.46	.33‡	.37‡	.36‡	.36‡	.26‡	.11‡
UEDIN-NMT	**.67‡**	.49	.50	–	.50	.43‡	.40‡	.36‡	.35‡	.35‡	.30‡	.14‡
ONLINE-G	**.67‡**	**.54⋆**	**.54**	.50	–	.46⋆	.40‡	.41‡	.39‡	.38‡	.33‡	.13‡
NYU-UMONTREAL	**.69‡**	**.55⋆**	**.54**	**.57‡**	**.54⋆**	–	.50	.42‡	.43‡	.43‡	.38‡	.16‡
JHU-PBMT	**.74‡**	**.67‡**	**.67‡**	**.60‡**	**.60‡**	.50	–	.43†	.46⋆	.40‡	.37‡	.20‡
LIMSI	**.69‡**	**.65‡**	**.63‡**	**.64‡**	**.59‡**	**.58‡**	**.57†**	–	**.51**	.45⋆	.40‡	.20‡
ONLINE-A	**.80‡**	**.68‡**	**.64‡**	**.65‡**	**.61‡**	**.57‡**	**.54⋆**	.49	–	.47	.42‡	.17‡
AFRL-MITLL-PHRASE	**.74‡**	**.69‡**	**.64‡**	**.65‡**	**.62‡**	**.57‡**	**.60‡**	**.55⋆**	**.53**	–	.41‡	.20‡
AFRL-MITLL-VERB-A	**.79‡**	**.72‡**	**.74‡**	**.70‡**	**.67‡**	**.62‡**	**.63‡**	**.60‡**	**.58‡**	**.59‡**	–	.25‡
ONLINE-F	**.93‡**	**.86‡**	**.89‡**	**.86‡**	**.87‡**	**.84‡**	**.80‡**	**.80‡**	**.83‡**	**.80‡**	**.75‡**	–
score	.78	.30	.26	.25	.20	.10	-.01	-.07	-.10	-.14	-.31	-.126
rank	1	2-4	2-5	2-5	3-5	6	7-8	7-10	8-10	9-10	11	12

Table 46: Head to head comparison, ignoring ties, for English-Russian systems

	AMU-UEDIN	ONLINE-G	NRC	ONLINE-B	UEDIN-NMT	ONLINE-A	AFRL-MITLL-PHRASE	AFRL-MITLL-CONTRA	PROMT-RULE-BASED	ONLINE-F
AMU-UEDIN	–	**.51**	.44†	.47	.41‡	.37‡	.38‡	.34‡	.35‡	.16‡
ONLINE-G	.49	–	.47	.44†	.41‡	.38‡	.41‡	.35‡	.36‡	.18‡
NRC	**.56**†	**.53**	–	.47	.45†	.40‡	.39‡	.38‡	.34‡	.19‡
ONLINE-B	**.53**	**.56**†	**.53**	–	.49	.44†	.42‡	.41‡	.36‡	.22‡
UEDIN-NMT	**.59**‡	**.59**‡	**.55**†	**.51**	–	.45†	.46⋆	.40‡	.44‡	.23‡
ONLINE-A	**.63**‡	**.62**‡	**.60**‡	**.56**†	**.55**†	–	.48	.47	.45†	.22‡
AFRL-MITLL-PHRASE	**.62**‡	**.59**‡	**.61**‡	**.58**‡	**.54**⋆	**.52**	–	.45†	.46†	.25‡
AFRL-MITLL-CONTRA	**.66**‡	**.65**‡	**.62**‡	**.59**‡	**.60**‡	**.53**	**.55**†	–	.50	.29‡
PROMT-RULE-BASED	**.65**‡	**.64**‡	**.66**‡	**.64**‡	**.56**‡	**.55**†	**.54**†	.50	–	.23‡
ONLINE-F	**.84**‡	**.82**‡	**.81**‡	**.78**‡	**.77**‡	**.78**‡	**.75**‡	**.71**‡	**.77**‡	–
score	.44	.42	.32	.25	.15	.03	.02	-.11	-.16	-.138
rank	1-2	1-3	2-4	3-5	4-5	6-7	6-7	8-9	8-9	10

Table 47: Head to head comparison, ignoring ties, for Russian-English systems

	ONLINE-B	ONLINE-G	ONLINE-A	TBTK-SYSCOMB	PROMT-SMT	YSDA	JHU-SYNTAX	JHU-PBMT	PARFDA
ONLINE-B	–	.44†	.45⋆	.35‡	.32‡	.31‡	.21‡	.20‡	.17‡
ONLINE-G	**.56**†	–	.47	.38‡	.36‡	.31‡	.19‡	.19‡	.19‡
ONLINE-A	**.55**⋆	**.53**	–	.41‡	.40‡	.35‡	.24‡	.15‡	.16‡
TBTK-SYSCOMB	**.65**‡	**.62**‡	**.59**‡	–	.47	.46	.26‡	.23‡	.23‡
PROMT-SMT	**.68**‡	**.64**‡	**.60**‡	**.53**	–	.46	.30‡	.29‡	.21‡
YSDA	**.69**‡	**.69**‡	**.65**‡	**.54**	**.54**	–	.32‡	.27‡	.26‡
JHU-SYNTAX	**.79**‡	**.81**‡	**.76**‡	**.74**‡	**.70**‡	**.68**‡	–	.47	.42⋆
JHU-PBMT	**.80**‡	**.81**‡	**.85**‡	**.77**‡	**.71**‡	**.73**‡	**.53**	–	.44
PARFDA	**.83**‡	**.81**‡	**.84**‡	**.77**‡	**.79**‡	**.74**‡	**.58**⋆	**.56**	–
score	.82	.65	.56	.21	.12	.00	-.67	-.76	-.93
rank	1-2	1-3	2-3	4-5	4-6	5-6	7-8	7-9	8-9

Table 48: Head to head comparison, ignoring ties, for Turkish-English systems

	ONLINE-G	ONLINE-B	ONLINE-A	YSDA	JHU-HLTCOE	TBTK-MORPH-HPB	CMU	JHU-PBMT	PARFDA
ONLINE-G	–	.45	.41†	.31‡	.26‡	.30‡	.25‡	.23‡	.16‡
ONLINE-B	**.55**	–	.46	.34‡	.29‡	.29‡	.30‡	.22‡	.18‡
ONLINE-A	**.59**†	**.54**	–	.42†	.38‡	.40‡	.29‡	.25‡	.25‡
YSDA	**.69**‡	**.66**‡	**.58**†	–	.43†	.44†	.40‡	.34‡	.31‡
JHU-HLTCOE	**.74**‡	**.71**‡	**.62**‡	**.57**†	–	.46	.45	.35‡	.35‡
TBTK-MORPH-HPB	**.70**‡	**.71**‡	**.60**‡	**.56**†	**.54**	–	.45⋆	.44†	.41‡
CMU	**.75**‡	**.70**‡	**.71**‡	**.60**‡	**.55**	**.55**⋆	–	.38‡	.42†
JHU-PBMT	**.77**‡	**.78**‡	**.75**‡	**.66**‡	**.65**‡	**.56**†	**.62**‡	–	.41†
PARFDA	**.84**‡	**.82**‡	**.75**‡	**.69**‡	**.65**‡	**.59**‡	**.58**†	**.59**†	–
score	.76	.61	.37	.05	-.12	-.19	-.29	-.54	-.66
rank	1-2	1-2	3	4	5-6	5-7	6-7	8-9	8-9

Table 49: Head to head comparison, ignoring ties, for English-Turkish systems