Findings of the WMT 2019 Shared Task on Automatic Post-Editing

Rajen Chatterjee(1), **Christian Federmann**(2) **Matteo Negri**(3), **Marco Turchi**(3)
(1) Apple Inc., Cupertino, CA, USA
(2) Microsoft Cloud+AI, Redmond, WA, USA
(3) Fondazione Bruno Kessler, Trento, Italy

Abstract

We present the results from the 5^{th} round of the WMT task on MT Automatic Post-Editing. The task consists in automatically correcting the output of a "black-box" machine translation system by learning from human corrections. Keeping the same general evaluation setting of the previous four rounds, this year we focused on two language pairs (English-German and English-Russian) and on domain-specific data (Information Technology). For both the language directions, MT outputs were produced by neural systems unknown to participants. Seven teams participated in the **English-German** task, with a total of 18 submitted runs. The evaluation, which was performed on the same test set used for the 2018 round, shows further progress in APE technology: 4 teams achieved better results than last year's winning system, with improvements up to -0.78 TER and +1.23 BLEU points over the baseline. Two teams participated in the **English-Russian** task submitting 2 runs each. On this new language direction, characterized by a higher quality of the original translations, the task proved to be particularly challenging. Indeed, none of the submitted runs improved the very high results of the strong system used to produce the initial translations (16.16 TER, 76.20 BLEU).

1 Introduction

MT Automatic Post-Editing (APE) is the task of automatically correcting errors in a machine-translated text. As pointed out by (Chatterjee et al., 2015), 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.

In its 5^{th} round, the APE shared task organized within the WMT Conference on Machine Translation kept the same overall evaluation setting of the previous four rounds. Specifically, the participating systems had to automatically correct the output of an unknown "black box" MT system by learning from human revisions of translations produced by the same system.

This year, the task focused on two language pairs (English-German and English-Russian) and, in continuity with the last three rounds, on data coming from the Information Technology domain. While in 2018 one of the proposed subtasks was still focusing on the correction of phrase-based MT output, this year only neural MT (NMT) output has been considered. However, this year's campaign allows both for a fair assessment of the progress in APE technology and for tests in more challenging conditions. On one side, reusing the same test English-German set used last year, the evaluation framework allows us for a direct comparison with the last year's outcomes at least on one language. On the other side, dealing with a

Proceedings of the Fourth Conference on Machine Translation (WMT), Volume 3: Shared Task Papers (Day 2) pages 13–30
Florence, Italy, August 1-2, 2019. ©2019 Association for Computational Linguistics

difficult language like Russian and only with high-quality NMT output, also this round presented participants with an increased level of difficulty with respect to the past.

Seven teams participated in the English-German task, submitting 18 runs in total. Two teams participated in the English-Russian task, submitting 2 runs each. Similar to last year, all the teams developed their systems based on neural technology, which confirms to be the state-of-the-art approach to APE. Only in one case, indeed, a participating team achieved its highest results (but with no improvement over the baseline) with a phrase-based APE system. In most of the cases, participants experimented with the Transformer architecture (Vaswani et al., 2017), either directly or by adapting it to the task (see Section 3). Another common trait of the submitted systems is the reliance on the consolidated multi-source approach (Zoph and Knight, 2016; Libovický et al., 2016), which is able to exploit information from both the MT output to be corrected and the corresponding source sentence. The third aspect common to all submissions is the exploitation of synthetic data, either those provided together with the task data (Negri et al., 2018; Junczys-Dowmunt and Grundkiewicz, 2016) or similar, domain-specific resources created *ad-hoc* by participants.

In the English-German task, the evaluation was performed on the same test set used in 2018, whose "gold" human post-edits were kept undisclosed to participants for the sake of future comparisons. Evaluating on the same benchmark allowed to observe further technology improvements over the past. Last year, the largest gain over the baseline (16.84 TER, 74.73 BLEU) was -0.38 TER (16.46) and +0.8 BLEU (75.53). This year, four teams achieved better results than last year's best submission. The top-ranked system achieved 16.06 TER (-0.78 with respect to the baseline) and 75.96 BLEU (+1.23). Most noticeably, the fact that the TER/BLEU differences between the top four primary submissions are not statistically significant indicates that the observed progress is not isolated.

The newly proposed English-Russian task represents a more challenging evaluation scenario, mainly due to the higher quality of the NMT output to be corrected. In this case, even the best submission (16.59 TER, 75.27 TER) was unable to beat the baseline (16.16 TER, 76.20 BLEU). These results confirm one of the main findings of previous rounds (Bojar et al., 2017; Chatterjee et al., 2018a): improving high-quality MT output remains the biggest challenge for APE. This motivates further research on precise and conservative solutions able to mimic human behaviour by performing only the minimum amount of edit operations needed.

2 Task description

In continuity with all the previous rounds of the APE task, 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.

2.1 Data

This year, the evaluation was performed on two language pairs, English-German and English-Russian. For both the subtasks, data were selected from the Information Technology (IT) domain. As emerged from the previous evaluations, the selected target domain is specific and repetitive enough to allow supervised systems to learn from the training set useful correction patterns that are also re-applicable to the test set.

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

- The source (SRC) is a tokenized English sentence;
- The target (TGT) is a tokenized German/Russian translation of the source, which was produced by a black-box system unknown to participants. For both the languages, translations were obtained from neural MT systems:[1] this implies that their overall quality is generally high, making the task harder compared to previous rounds, which

[1] For **English-German**, the NMT system was trained with generic and in-domain parallel training data using the attentional encoder-decoder architecture (Bahdanau et al., 2014) implemented in the Nematus toolkit (Sennrich et al., 2017). We used byte-pair encoding (Sennrich et al., 2016) for vocabulary reduction, mini-batches of 100, word embeddings of 500 dimensions, and gated recurrent unit layers of 1,024 units. Optimization was done using Adam and by re-shuffling the training set at each epoch. For **English-Russian**, the NMT system used was the Microsoft Translator production system, which was trained with both generic and in-domain parallel training data.

	Number of instances			
	Training	Development	Test	Additional Resources
English-German	13,442	1,000	1,023	eSCAPE-PBSMT: 7,258,533 eSCAPE-NMT: 7,258,533 Artificial: 4.5M
English-Russian	15,089	1,000	1,023	eSCAPE-NMT: 7.7 M

Table 1: Data statistics.

focused only (until 2017) or also (as in 2018) on the correction of the output of phrase-based systems.

- The human post-edit (PE) is a manually-revised version of the target, which was produced by professional translators.

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

For the **English-German** subtask, the same in-domain data[2] collected for last year's round of the task have been used. The *training* and *development* set respectively contain 13,442 and 1,000 triplets, while the *test* set consists of 1,023 instances. Participants were also provided with two additional training resources, which were widely used in last year's round. One (called "Artificial" in Table 1) is the corpus of 4.5 million artificially-generated post-editing triplets described in (Junczys-Dowmunt and Grundkiewicz, 2016). The other resource is the English-German section of the eSCAPE corpus (Negri et al., 2018). It comprises 14.5 million instances, which were artificially generated both via phrase-based and neural translation (7.25 millions each) of the same source sentences.

For the **English-Russian** subtask, Microsoft Office localization data have been used. This material, which mainly consists of short segments (menu commands, short messages, etc.), is shared with the English-Russian Quality Estimation shared task.[3] The *training* and *development* set respectively contain 15,089 and 1,000 triplets, while the *test* set comprises 1,023 instances. For this language pair, the eSCAPE corpus has been extended to provide participants with additional training material.[4]

Table 1 provides basic statistics about the data of the two subtasks.

2.1.1 Complexity indicators: repetition rate

Table 2 provides a view of the data from a task difficulty standpoint. For each dataset released in the five rounds of the APE task, it shows the repetition rate of SRC, TGT and PE elements, as well as the TER (Snover et al., 2006) and the BLEU score (Papineni et al., 2002) of the TGT elements (i.e. the original target translations).

The 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 values indicate a higher text repetitiveness and, as discussed in (Bojar et al., 2016; Bojar et al., 2017; Chatterjee et al., 2018a), suggest a higher chance of learning from the training set correction patterns that are applicable also to the test set. In the previous rounds of the task, we considered the large differences in repetitiveness across the datasets as a possible explanation for the variable gains over the baseline obtained by participants. In this perspective, the low system performance observed in the APE15 task and in the APE17 German-English subtask was in part ascribed to the low repetition rate in the data. In contrast, much higher repetition rates in the data likely contributed to facilitate the problem in the APE16 task and in the APE17 English-German subtask, in which most of the participants achieved significant gains over the baseline. Although in both the APE18 subtasks the repetition rate values were relatively high, evaluation results shown that the influence of data repetitiveness on final APE performance is marginal. Indeed, while in the last year's PBSMT subtask the improvements over the baseline were impressive (-

[2] Released by the European Project QT21 (Specia et al., 2017).

[3] `http://www.statmt.org/wmt19/qe-task.html`

[4] This newly released artificial dataset and a short description of the methodology adopted for its creation can be found at `http://hltshare.fbk.eu/QT21/eSCAPE.html`.

	2015	2016	2017	2017	2018	2018	2019	2019
Language	En-Es	En-De	En-De	De-En	En-De	En-De	En-De	En-Ru
Domain	News	IT	IT	Medical	IT	IT	IT	IT
MT type	PBSMT	PBSMT	PBSMT	PBSMT	PBSMT	NMT	NMT	NMT
Repet. Rate SRC	2.905	6.616	7.216	5.225	7.139	7.111	7.111	18.25
Repet. Rate TGT	3.312	8.845	9.531	6.841	9.471	9.441	9.441	14.78
Repet. Rate PE	3.085	8.245	8.946	6.293	8.934	8.941	8.941	13.24
Baseline TER	23.84	24.76	24.48	15.55	24.24	16.84	16.84	16.16
Baseline BLEU	n/a	62.11	62.49	79.54	62.99	74.73	74.73	76.20

Table 2: Basic information about the APE shared task data released since 2015: languages, domain, type of MT technology, repetition rate and initial translation quality (TER/BLEU of TGT). Grey columns refer to data covering different language pairs and domains with respect to this year's evaluation round.

6.24 TER, +9.53 BLEU points), in the NMT subtask (whose data were reused this year) the quality gains were considerably smaller (-0.38 TER and +0.8 BLEU points). As discussed in Section 4.1, also this year we observe a similar situation: especially for English-Russian, the high repetition rate values reported in Table 2, which are the highest ones across all the APE data released so far, are not enough to determine quality improvements comparable to previous rounds. This suggests that, although it used to play an important role when dealing with lower quality MT output in the first rounds of the APE task, text repetitiveness has less influence on final performance compared to other complexity indicators.

2.1.2 Complexity indicators: MT quality

Indeed, another important aspect that determines the difficulty of APE is the initial quality of the MT output to be corrected. This can be measured by computing the TER (↓) and BLEU (↑) scores (last two rows in Table 2) using the human post-edits as reference.

As discussed in (Bojar et al., 2017; Chatterjee et al., 2018a), numeric evidence of a higher quality of the original translations can indicate a smaller room for improvement for APE systems (having, at the same time, less to learn during training and less to correct at test stage). On one side, indeed, training on good (or near-perfect) automatic translations can drastically reduce the number of learned correction patterns. On the other side, testing on similarly good translations can drastically reduce the number of corrections required and the applicability of the learned patterns, thus making the task more difficult. As observed in the previous APE evaluation rounds, there is a noticeable correlation between translation quality and systems' performance. In 2016 and 2017, on English-German data featuring a similar level of quality (24.76/24.48 TER, 62.11/62.49 BLEU), the top systems achieved significant improvements over the baseline (-3.24 TER and +5.54 BLEU in 2016, -4.88 TER and +7.58 BLEU in 2017). In 2017, on higher quality German-English data (15.55 TER, 79.54 BLEU), the observed gains were much smaller (-0.26 TER, +0.28 BLEU). In 2018, the correction of English-German translations produced by a phrase-based system (24.24 TER, 62.99 BLEU) yielded much larger gains (up to -6.24 TER and +9.53 BLEU) compared to the correction of higher-quality neural translations (16.84 TER, 74.73 BLEU), which resulted in TER/BLEU variations of less than 1.00 point. As discussed in Section 4, also this year's results confirm the strict correlation between the quality of the initial translations and the actual potential of APE.

2.1.3 Complexity indicators: TER distribution

Further indications about the difficulty of the two subtasks are provided by Figures 1 and 2, which plot the TER distribution for the items in the two test sets. As shown in Figure 1, the distribution for English-German is quite skewed towards low TER values, with almost 50% of the test test items having a TER between 0 and 10 that indicates their very high quality (in other terms, they require few edits). In particular, the proportion of "perfect" test instances having TER=0 (i.e. items that should not be modified by the APE systems) is quite high (25.2% of the total).[5] For these test

[5] This value is considerably lower than the proportion observed in the challenging APE17 German-English test set (45.0%) but still a considerably higher value compared to "easier" test sets released for other rounds of the task.

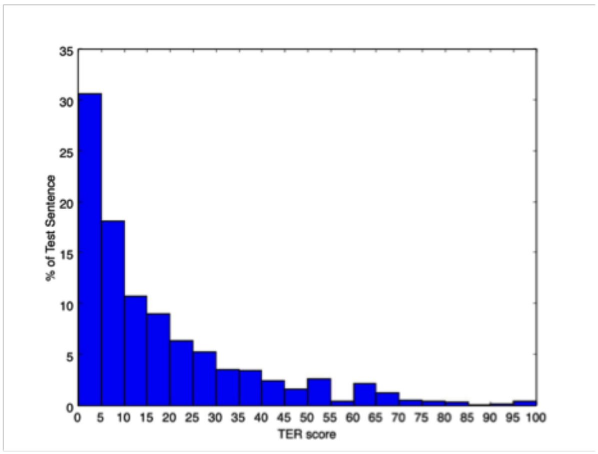


Figure 1: TER distribution in the **English-German** test set

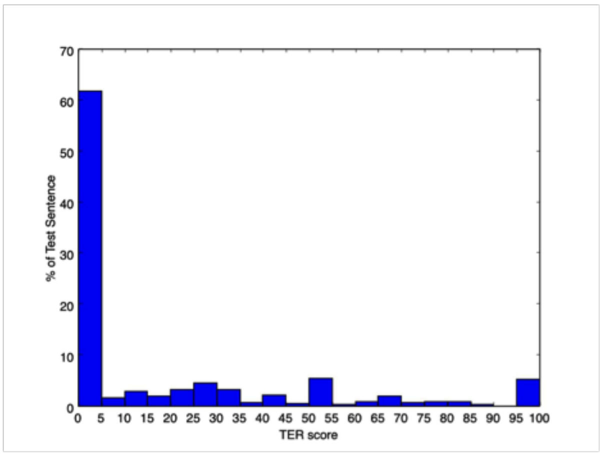


Figure 2: TER distribution in the **English-Russian** test set

items, any correction made by the APE systems will be treated as unnecessary and penalized by automatic evaluation metrics. This problem calls for conservative and precise systems able to properly fix errors only in the remaining test items, leaving the "perfect" ones unmodified.

Data skewedness is exacerbated in the English-Russian test set, in which 63.5% of the instances have a TER between 0 and 10 (in particular, 61.4% of them are perfect translations). Together with the high BLEU score, this contributes to make the English-Russian task considerably more difficult compared to the English-German one (as well as compared to most of the APE test sets released so far).

As discussed in Section 4, also this year's evaluation results confirm the strict correlation between the quality of the initial translations and the actual potential of APE.

2.2 Evaluation metrics

System performance was evaluated both by means of automatic metrics and manually. Automatic metrics were used to compute the distance between *automatic* and *human* post-edits of the machine-translated sentences present in the test sets. To this aim, TER and BLEU (case-sensitive) were respectively used as primary and secondary evaluation metrics. Systems were ranked based on the average TER calculated on the test set by using the TERcom[6] software: lower average TER scores correspond to higher ranks. BLEU was computed using the multi-bleu.perl package[7] available in MOSES.

Manual evaluation was conducted via source-based direct human assessment (Graham et al., 2013; Cettolo et al., 2017; Bojar et al., 2018) as implemented by Appraise (Federmann, 2012). Details are discussed in Section 6.

2.3 Baseline

In continuity with the previous rounds, the official baseline results were the TER and BLEU scores calculated by comparing the raw MT output with the human post-edits. In practice, the baseline APE system is a "*do-nothing*" system that leaves all the test targets unmodified. Baseline results, the same shown in Table 2, are also reported in Tables 4 and 5 for comparison with participants' submissions.[8]

For each submitted run, the statistical significance of performance differences with respect to the baseline was calculated with the bootstrap test (Koehn, 2004).

3 Participants

Seven teams submitted a total of 18 runs for the English-German subtask. Two of them participated also in the English-Russian subtask by sub-

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

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

[8] In addition to the *do-nothing* baseline, in the first three rounds of the task we also compared systems' performance with a re-implementation of the phrase-based approach firstly proposed by Simard et al. (2007), which represented the common backbone of APE systems before the spread of neural solutions. As shown in (Bojar et al., 2016; Bojar et al., 2017), the steady progress of neural APE technology has made the phrase-based solution not competitive with current methods reducing the importance of having it as an additional term of comparison. In 2018, we hence opted for considering only one baseline.

ID	Participating team
ADAPT_DCU	ADAPT Centre & Dublin City University, Ireland (Shterionov et al., 2019)
FBK	Fondazione Bruno Kessler, Italy (Tebbifakhr et al., 2019)
POSTECH	Pohang University of Science and Technology, South Korea (Lee et al., 2019)
UDS	Saarland University, Germany (Xu et al., 2019)
UNBABEL	Unbabel, Portugal (Lopes et al., 2019)
USAAR_DFKI	Saarland University & German Research Center for Artificial Intelligence, Germany (Pal et al., 2019)
IC_USFD	Imperial College London & University of Sheffield, United Kingdom

Table 3: Participants in the WMT19 Automatic Post-Editing task.

mitting 2 runs each. Participants are listed in Table 3, and a short description of their systems is provided in the following.

ADAPT Centre & Dublin City University. The ADAPT_DCU team participated in both the subtasks proposed this year. Their submissions pursue two main objectives, namely: *i)* investigating the effect of adding extra information in the form of prefix tokens in a neural APE system; and *ii)* assessing whether an SMT-based approach can be effective for post-editing NMT output. The neural APE system exploits a multi-source approach based on Marian-NMT.[9] Training data were augmented with two types of extra context tokens that identify partitions of the training set that may be relevant to guide system's behaviour (i.e. to identify features in the dataset with a very close relation to the editing patterns the system is supposed to learn). Such partitions are based on sentence length and topic information. Hence, the prepended tokens respectively state the data partition based on the number of source tokens and the topic induced via LDA clustering (Blei et al., 2003). The statistical APE models, which are based on Moses (Koehn et al., 2007), were trained to explore the idea of interleaving different MT technologies to improve NMT output quality. All the models are built by taking advantage of both the released training material and the provided artificial data (Negri et al., 2018; Junczys-Dowmunt and Grundkiewicz, 2016).

Fondazione Bruno Kessler. Also FBK participated in both the subtasks. Their submissions focus on mitigating the "over-correction" problem in APE, that is the systems' tendency to rephrase and correct MT output that is already acceptable, thus producing translations that will be penalized by evaluation against human post-edits. Following (Chatterjee et al., 2018b), the underlying idea is that over-correction can be prevented by informing the system about the predicted quality of the MT output or, in other terms, the expected amount of corrections needed. The proposed solution is based on prepending a special token to the source text and the MT output, so to indicate the required amount of post-editing. Three different tokens are used, namely "no post-edit" (no edits are required), "light post-edit" (minimal edits are required), and "heavy post-edit" (a large number of edits are required. At training time, the instances are labelled based on the TER computed between the MT output and its post-edited version, with the boundary between light and heavy post-edit set to TER=0.4 based on the findings reported in (Turchi et al., 2013; Turchi et al., 2014). At test time, tokens are predicted with two approaches. One is based on a classifier obtained by fine-tuning BERT (Devlin et al., 2018) on the in-domain data. The other approach exploits a retrieval-based method similar to (Farajian et al., 2017): given a query containing the source and the MT output to be post-edited, it: *i)* retrieves similar triplets from the training data, *ii)* ranks them based on the sentence level BLEU score between the MT output and the post-edit, and *iii)* creates the token based on the TER computed between the MT output and the post-edit of the most similar triplet. The backbone architecture is the multi-source extension of Transformer (Vaswani et al., 2017) described in (Tebbifakhr et al., 2018), which is trained both on the task data and on the available artificial corpora.

Pohang University of Science and Technology. POSTECH's system (English-German subtask) is a multi-source model that extends the Transformer implementation of the OpenNMT-py (Klein et al., 2017) library. It includes: *i)* a joint encoder that is able to generate joint representations reflecting the relationship between two input sources (SRC, TGT) with optional future masking to mimic the general decoding process of machine translation systems, and *ii)* two types of multi-source attention layers in the decoder that computes the atten-

[9] https://marian-nmt.github.io/

tion between the decoder state and the two outputs of the encoder. Therefore, four different model variants were suggested in terms of the existence of the encoder future mask and the type of the multi-source attention layer in the decoder. The eSCAPE corpus (Negri et al., 2018) was filtered to contain similar statistics as the official training dataset. During training, various teacher-forcing ratios were adjusted to alleviate the exposure bias problem. After training four variants with various teacher-forcing ratios, the final submissions were obtained from an ensemble of models. These are: 1) the primary submission that ensembles the variants with the two best TER scores in each architecture, 2) the contrastive submission that ensembles the variants with the best TER scores in each architecture, 3) the contrastive submission that ensembles two variants from each architecture, achieving the best TER and BLEU, respectively.

Saarland University. UdS's participation (English-German subtask) is based on a multi-source Transformer model for context-level machine translation (Zhang et al., 2018) implemented with the Neutron implementation (Xu and Liu, 2019) for the Transformer translation model (Vaswani et al., 2017). To improve the robustness of the training, and inspired by (Cheng et al., 2018), the APE task is jointly trained with the de-noising encoder task, which adds noises distribution directly to the post-editing results' embedding as machine translation outputs and tries to recover the original post-editing results. Both Gaussian noise and uniform noise were tried for the de-noising encoder task. The synthetic eSCAPE corpus (Negri et al., 2018) was also used for the training. Contrastive submissions were generated with the best averaged models of 5 adjacent checkpoints of 2 kinds of noise, and the primary submission is obtained with the ensemble of 5 models (4 averaged models + 1 model saved for every training epoch).

Unbabel. Following (Correia and Martins, 2019), Unbabel's submission (English-German subtask) adapts BERT (Devlin et al., 2018) to the APE task with an encoder-decoder framework. The system consists in a BERT encoder initialised with the pretrained model's weights and a BERT decoder initialised analogously, where the multi-head context attention is initialised with the self-attention weights. Additionally, source embeddings, target embeddings and projection layer (Press and Wolf, 2017) are shared, as well as the self-attention weights of the encoder and decoder. The system exploits BERT training schedule with streams A and B: the encoder receives as input both the source and the MT output separated by the special symbol "[SEP]", assigning to the first "A" segment embeddings and to the latter "B" segment embeddings. Regarding the BERT decoder, they use just the post-edit with "B" segment embeddings. In addition, as the NMT system has a strong in-domain performance, a conservativeness factor to avoid over-correction is explored. Similarly to (Junczys-Dowmunt and Grundkiewicz, 2016), a penalty is added during beam decoding (logits or log probabilities) to constrain the decoding to be as close as possible to the input – both the source and the MT output are considered, which allows to copy from the source – in order to avoid over edition of the MT segment. This penalty is tuned over the development set. In addition to the shared task in-domain data, system training exploits a variant of the eSCAPE corpus built on a closer in-domain parallel corpus (IT domain) provided by the Quality Estimation shared task.

Saarland University & German Research Center for Artificial Intelligence. USAAR_DFKI's participation (English-German subtask) is based on a multi-encoder adaptation of the Transformer architecture. The system consists in: *i)* a Transformer encoder block for the source sentence, followed by *ii)* a Transformer decoder block, but without masking, for self-attention on the MT segment, which effectively acts as second encoder combining source and MT output, and *iii)* feeds this representation into a final decoder block generating the post-edit. The intuition behind the proposed architecture is to generate better representations via both self- and cross- attention and to further facilitate the learning capacity of the feed-forward layer in the decoder block. Also in this case, model training takes advantage of the eSCAPE synthetic data (Negri et al., 2018).

University of Sheffield & Imperial College London. IC_USFD's submission (English-German subtask) is based on the dual-source Transformer model (Junczys-Dowmunt and Grundkiewicz, 2018), which was re-implemented in the

Tensor2Tensor (Vaswani et al., 2017) toolkit. The model was enriched with a copying mechanism that prevents unnecessary corrections. In addition to the main training data, the primary submission uses the EN-DE eSCAPE data (Negri et al., 2018). The contrastive submission uses data triplets where source and target are genuine data, and MT is a modified target (200K). This modified target mimics MT by simulating errors in the task training data. Sentences where error simulation is possible are selected from in-domain corpora (eSCAPE, as well as the in-domain data released with the WMT18 Quality Estimation task).

4 Results

Participants' results are shown in Tables 4 (English-German) and 5 (English-Russian). The submitted runs are ranked based on the average TER (case-sensitive) computed using human post-edits of the MT segments as reference, which is the APE task primary evaluation metric ("*TER (pe)*"). The two tables also report the BLEU score computed using human post-edits ("*BLEU (pe)*" column), which represents our secondary evaluation metric. These results are discussed in Section 4.1.

Table 4 includes four additional columns, which show the TER/BLEU scores computed using external references ("*TER (ref)*" and "*BLEU (ref)*") as well as the multi-reference TER/BLEU scores computed using human post-edits and external references ("*TER (pe+ref)*" and "*BLEU (pe+ref)*"). In Section 4.2, these figures are respectively used to discuss systems' capability to reflect the post-editing style of the training data and their tendency to produce unnecessary corrections of acceptable MT output. Since external references are available only for German, this analysis was not feasible for the English-Russian task.

4.1 Automatic metrics computed using human post-edits

Different from the past, this year the primary ("*TER (pe)*") and secondary evaluation metric ("*BLEU (pe)*") produce slightly different rankings.[10] For English-German, system results are quite close to each other, up to the point that *i)* TER differences between the top eight submissions are not statistically significant and *ii)* all the submissions with a TER score equal or lower than the baseline are concentrated in a performance interval of less than 0.8 TER points and less than 1.2 BLEU points. This compression can contribute to explain the ranking differences, especially at higher ranks where discriminating between strong systems with almost identical performance is particularly difficult. However, for the sake of future analysis or alternative views of this year's outcomes, it's worth remarking that the 2^{nd}, 3^{rd} and 5^{th} runs in terms of TER (all by the same team –POSTECH) respectively represent the top three submissions in terms of BLEU.

For English-Russian, the distance between the top and the worst submissions is larger, but also in this case the BLEU-based ranking is not identical to the TER-based one. Though with a negligible margin, the worst run in terms of TER would rank 2^{nd} in terms of BLEU.

English-German subtask. In order to measure the progress with respect to last year's round of the APE task, for this language pair the evaluation has been performed with the same data used for the NMT subtask in 2018. Last year, the majority of the participants' scores fell in a range of less than one TER/BLEU point improvement over the *do-nothing* baseline (16.84 TER, 74.73 BLEU), being 16.46 TER (-0.38) and 75.53 BLEU (+0.8) the scores and the corresponding quality gains achieved by the top submission. This year, eight submissions achieved a TER reduction larger than 0.4 points and a BLEU increase of more than 0.9 points. The top submission, in particular, obtained improvements up to -0.78 TER and +1.23 BLEU points over the baseline. Although correcting the output of a neural MT system still proves to be quite hard, we take the fact that 4 teams achieved better results than last year's winning system as an indicator of technology advancements.

English-Russian subtask. This subtask proved to be more challenging compared to the English-German subtask. Final results are indeed much worse: none of the four runs submitted by the two participating teams was able to beat the *do-nothing* baseline (16.16 TER, 76.2 BLEU). Even for the top submission (16.59 TER, 75.27 BLEU), results' difference with respect to the baseline is statistically significant. One possible cause of the higher difficulty of the English-Russian subtask is the fact that dealing with a morphology-rich lan-

[10] The correlation between the ranks obtained by the two metrics is 0.97 for the English-German subtask and 0.7 for the English-Russian subtask.

ID	TER (pe)	BLEU (pe)	TER (ref)	BLEU (ref)	TER (pe+ref)	BLEU (pe+ref)
UNBABEL Primary	16.06*	75.96	41.66	44.95	15.58	78.1
POSTECH Primary	16.11*	76.22	42.04	44.57	15.68	78.08
POSTECH Contrastive (var2Ens8)	16.13*	76.21	42.09	44.53	15.73	78.05
USAAR_DFKI Primary	16.15*	75.75	41.84	44.65	15.69	77.84
POSTECH Contrastive (top1Ens4)	16.17*	76.15	42.09	44.52	15.74	78.01
UNBABEL Contrastive (2)	16.21*	75.7	41.59	45.08	15.72	77.98
UNBABEL Contrastive (1)	16.24*	75.7	41.62	45.01	15.76	77.97
FBK Primary	16.37*	75.71	42.18	44.39	15.90	77.54
FBK Contrastive	16.61†	75.28	42.12	44.49	16.1	77.43
UDS Primary	16.77†	75.03	42.64	43.78	16.34	76.83
IC_USFD Contrastive	16.78†	74.88	42.45	44.01	16.31	76.82
UDS Contrastive (Gaus)	16.79†	75.03	42.55	44.0	16.33	76.87
UDS Contrastive (Uni)	16.80†	75.03	42.66	43.79	16.37	76.85
IC_USFD Primary	16.84†	74.8†	42.58	43.86	16.41	76.68
Baseline	16.84	74.73	42.24	44.2	16.27	76.83
ADAPT_DCU Contrastive (SMT)	17.07	74.3	42.40	44.14	16.54	76.36
ADAPT_DCU Primary	17.29	74.29	42.41	44.09	16.81	76.51
USAAR_DFKI Contrastive	17.31	73.97	42.45	43.71	16.87	76.06
ADAPT_DCU Contrastive (LEN)	17.41	74.01	42.44	44.01	16.91	76.2

Table 4: Results for the WMT19 APE **English-German subtask** – average TER (↓), BLEU score (↑). The symbol "*" indicates results differences between runs that are not statistically significant. The symbol "†" indicates a difference from the MT baseline that is not statistically significant.

ID	TER (pe)	BLEU (pe)
Baseline	16.16	76.2
ADAPT_DCU Contrastive	16.59	75.27
ADAPT_DCU Primary	18.31	72.9
FBK Primary	19.34	72.42
FBK Contrastive	19.48	72.91

Table 5: Results for the WMT19 APE **English-Russian subtask** – average TER (↓), BLEU score (↑).

guage like Russian is problematic not only for MT but also from the APE standpoint. Under similar data conditions (the training sets of the two subtasks differ by ∼1,650 instances), the training set of a morphology-rich language is likely to be more sparse compared to other languages. The other possible explanation lies in the higher quality of the original translations (our second complexity indicator discussed in Section 2.1.2), which reduces the room for improvement with APE and, at the same time, increases the possibility to damage MT output that is already correct. From the MT quality point of view, according to the baseline results shown in Table 2, the English-Russian dataset used for this year's campaign is the second most difficult benchmark released in five rounds of the APE task. Also the TER distribution of the test set instances (our third complexity indicator discussed in Section 2.1.3) indicates the higher difficulty of the task, which is characterized by the highest number of perfect translations across the five rounds of the APE shared task (61.4%). In terms of repetition rate, as observed in Section 2.1.1, English-Russian data considerably differ from those released for the previous rounds of the task. The much larger values shown in Table 2 are not surprising considering that this material is drawn from Microsoft Office localization data that mainly consist of short segments (e.g. menu commands), which are likely produced based on standardized guidelines. However, also this year text repetitiveness seems to have a smaller influence on final performance compared to quality of the initial translations. Besides all these elements, the higher difficulty of the English-Russian subtask is also indirectly suggested by the low number of participants. Likely, poor results observed on the development set during system development (i.e. the difficulty to beat the *do-nothing* baseline) discouraged other potential participants.

4.2 Automatic metrics computed using external references

By learning from (SRC, TGT, PE) triplets, APE systems' goal is to perform a "monolingual translation" from raw MT output into its correct version. In this translation process, the same sentence can be corrected in many possible ways that make the space of possible valid outputs potentially very large. Ideally, from this space, APE systems should select solutions that reflect as much as possible the post-editing style of the training data (in real-use settings, this can be the style/lexicon of specific users, companies, etc.). However, nothing prevents to end up with outputs that partially satisfy this constraint. In light of these considerations, TER and BLEU scores computed using human post-edits as reference represent a reliable measure of quality but:

1. They provide us with partial information on how systems' output reflects the post-editing style of the training data;

2. They are not informative at all about the amount of valid corrections that are not present in the human post-edits.

In order to shed light on these aspects, in previous rounds of the task, further analysis was performed by taking advantage of reference translations. In continuity with the past, in Sections 4.2.1 and 4.2.2 we re-propose this analysis for the English-German subtask, the only one for which external references are available.

4.2.1 Output style

To gain further insights on point 1. (i.e. system's capability to learn the post-editing style of the training data), the "*TER (ref)*" and "*BLEU (ref)*" columns in Table 4 show the TER and BLEU scores computed against independent reference translations. The rational behind their computation is that differences in TER/BLEU(pe) and TER/BLEU(ref) can be used as indicators of the "direction" taken by the trained models (i.e. either towards humans' post-editing style or towards a generic improvement of the MT output). Since independent references are usually very different from conservative human post-edits of the same TGT sentences, all the TER/BLEU scores measured using independent references are expected to be worse. However, if our hypothesis holds true, visible differences in the baseline improvements measured with TER/BLEU(pe) and TER/BLEU(ref) should indicate system's ability to model the post-editing style of the training data. In particular, larger gains measured with TER/BLEU(pe) will be associated to this desired ability.

As can be seen in Table 4, systems' results on English-German show this tendency. Looking at the improvements over the baseline, those measured by computing TER and BLEU scores against human post-edits (i.e. TER/BLEU(pe)) are often larger than those computed against independent references (i.e. TER/BLEU(ref)). In terms of TER, this holds true for most of the submitted runs, with the best system that shows a difference of 0.2 TER points in the gains over the baseline computed with TER(pe) (-0.78) and those computed with TER(ref) (-0.58). On average, for the runs achieving improvements over the baseline, the difference in the gains over the baseline computed with TER(pe) and TER(ref) is respectively -0.41 and -0.08. In terms of BLEU, the differences are more visible. The best system improves over the baseline by 1.23 points with BLEU(pe) and 0.75 points with BLEU(ref), while the average difference in the gains over the baseline is 0.8 with BLEU(pe) and 0.2 with BLEU (ref). The larger (0.32/0.6) average improvements over the baseline observed with TER/BLEU computations against human post-edits can be explained by systems' tendency to reflect the post-editing style of the training data.

4.2.2 Over-corrections

To shed light on point 2. (i.e. system's tendency to produce unnecessary corrections of acceptable MT output), the "*TER (pe+ref)*" and "*BLEU (pe+ref)*" columns in Table 4 show the multi-reference TER and BLEU scores computed against post-edits and independent references. The rational behind their computation is that differences in TER/BLEU(pe) and TER/BLEU(pe+ref) can be used to analyze the quality of the unnecessary corrections performed by the systems (or, in other words, to study the impact of systems' tendency towards "over-correction”). APE corrections of a given MT output can indeed be of different types, namely: *i)* correct edits of a wrong passage, *ii)* wrong edits of a wrong passage, *iii)* correct edits of a correct passage and *iv)* wrong edits of a correct passage. TER/BLEU scores computed against human post-edits work reasonably well in capturing cases *i)-ii)* by matching APE systems' output with human post-edits: for wrong MT output passages (i.e. those changed by the post-editor), they inform us about the general quality of automatic corrections (i.e. how close they are to the post-editor's actions). Cases *iii)-iv)*, in contrast, are more problematic since any change performed by the system to a correct passage (i.e. those that were not changed by the post-editor) will always be penalized by automatic comparisons with human post-edits. Although discriminating between the two types of unnecessary corrections is hard, we hypothesize that a comparison between TER/BLEU(pe) and TER/BLEU(pe+ref) can be used as a proxy to quantify those belonging to type *iii)*. In general, due to the possibility to match more and longer n-grams in a multi-reference setting, TER/BLEU(pe+ref) scores are expected to be higher than TER/BLEU(pe) scores. However, if our hypothesis holds true, visible differences in the increase observed for the baseline and for the systems should indicate systems' tendency to produce acceptable over-corrections (type *iii)*). In particular, larger gains observed for the APE systems will be associated to their over-correction tendency towards potentially acceptable edits that should not be penalized by automatic evaluation metrics.

As expected, Table 4 shows that, on English-German data, multi-reference evaluation against post edits and external references (TER/BLEU(pe+ref)) yields better results compared to single reference evaluation with post-edits only (TER/BLEU(pe)). The variations of the *do-nothing* baseline are -0.57 TER (from 16.84 to 16.27) and 2.1 BLEU (from 74.73 to 76.83) points. In contrast, systems' scores vary by -0.46 TER and +2.01 BLEU points on average. In comparison with the larger variation observed for the baseline, this indicates that, for most of the submissions, the multi-reference evaluation does not indicate a tendency to produce unnecessary but acceptable corrections. On a positive note, while last year this was true for all the systems, this year four runs perform slightly better than the baseline in terms of BLEU(pe+ref). Though minimal, these differences suggest that a certain amount of corrections made by the top systems still represent acceptable modifications of the original translations.

5 System/performance analysis

As a complement to global TER/BLEU scores, also this year we performed a more fine-grained analysis of the changes made by each system to the test instances.

5.1 Macro indicators: modified, improved and deteriorated sentences

Tables 6 and 7 show the number of modified, improved and deteriorated sentences, respectively for the English-German and the English-Russian subtasks. It's worth noting that, as in the previous rounds and in both the settings, the number of sentences modified by each system is higher than the sum of the improved and the deteriorated ones. This difference is represented by modified sentences for which the corrections do not yield TER variations. This grey area, for which quality improvement/degradation can not be automatically assessed, contributes to motivate the human evaluation discussed in Section 6.

English-German subtask. As shown in table 6, the amount of sentences modified by the participating systems varies considerably. With values ranging from 4.01% to 39.1%, the average proportion of modifications (23.53%) is lower compared to last year (32.7%). Considering that about 25.2% (i.e. 257) of the test instances are to be considered as “perfect” (see Figure1), also this year the reported numbers are, for most of the submissions, far below the target percentage of modifications (74.8%). Overall, system's aggressiveness does not correlate with the final ranking: among both the top ranked systems and those with lower performance, large differences in the proportion of modified sentences can be observed. Indeed, as expected, what makes the difference is system's precision (i.e. the proportion of improved sentences out of the total amount of modified test items). Overall, the average precision is 45.92%, which represents a significant increase from last year's value (34.3%). While in 2018 none of the systems showed a precision higher than 50.0%, this year seven runs are above this value. As a

Systems	Modified	Improved	Deteriorated
UNBABEL Primary	366 (35.78%)	187 (51.09%)	110 (30.05%)
POSTECH Primary	207 (20.23%)	127 (61.35%)	41 (19.81%)
POSTECH Contrastive (var2Ens8)	210 (20.53%)	125 (59.52%)	45 (21.43%)
USAAR_DFKI Primary	301 (29.42%)	157 (52.16%)	83 (27.57%)
POSTECH Contrastive (top1Ens4)	213 (20.82%)	125 (58.69%)	47 (22.07%)
UNBABEL Contrastive (2)	400 (39.1%)	202 (50.50%)	121 (30.25%)
UNBABEL Contrastive (1)	393 (38.42%)	195 (49.62%)	117 (29.77%)
FBK Primary	200 (19.55%)	115 (57.50%)	50 (25.00%)
FBK Contrastive	363 (35.48%)	164 (45.18%)	131 (36.09%)
UDS Primary	96 (9.38%)	42 (43.75%)	36 (37.50%)
IC_USFD Contrastive	41 (4.01%)	21 (51.22%)	16 (39.02%)
UDS Contrastive (Gaus)	125 (12.22%)	54 (43.20%)	51 (40.80%)
UDS Contrastive (Uni)	112 (10.95%)	49 (43.75%)	41 (36.61%)
IC_USFD Primary	72 (7.04%)	29 (40.28%)	35 (48.61%)
ADAPT_DCU Contrastive (SMT)	120 (11.73%)	29 (24.17%)	61 (50.83%)
ADAPT_DCU Primary	368 (35.97%)	116 (31.52%)	169 (45.92%)
USAAR_DFKI Contrastive	391 (38.22%)	135 (34.53%)	168 (42.97%)
ADAPT_DCU Contrastive (LEN)	354 (34.60%)	101 (28.53%)	169 (47.74%)

Table 6: Number of test sentences modified, improved and deteriorated by each run submitted to the **English-German subtask**.

Systems	Modified	Improved	Deteriorated
ADAPT_DCU Contrastive	92 (8.99%)	17 (18.48%)	49 (53.26%)
ADAPT_DCU Primary	245 (23.95%)	57 (23.27%)	130 (53.06%)
FBK Primary	147 (14.37%)	49 (33.33%)	67 (45.58%)
FBK Contrastive	26 (2.54%)	5 (19.23%)	18 (69.23%)

Table 7: Number of test sentences modified, improved and deteriorated by each run submitted to the **English-Russian subtask**.

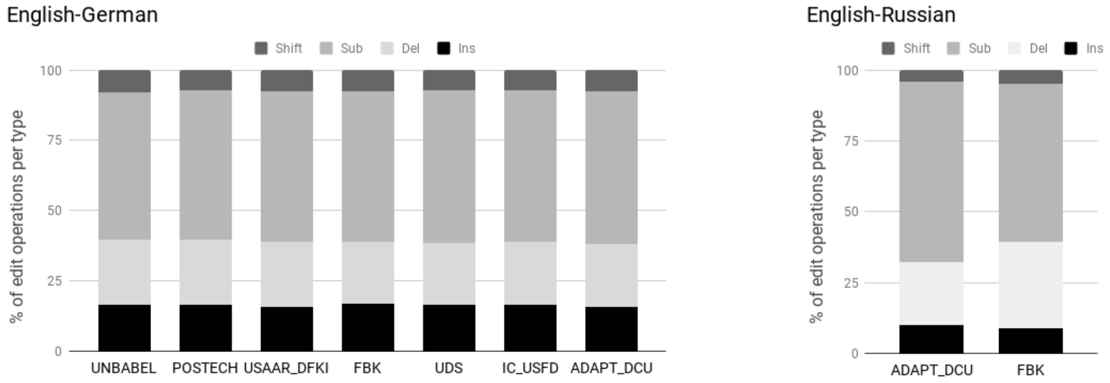
consequence, the percentage of deteriorated sentences out of the total amount of modified test items shows a significant drop. On average, a quality decrease is observed for 35.11% of the test items, while last year the average was 47.85%.

English-Russian subtask. As shown in table 7, also in this subtask the amount of sentences modified by the submitted systems varies considerably and does not correlate with systems' ranking. On average, the proportion of modifications is 12.46% (much less compared to the English-German subtask). With values ranging from 2.54% to 23.95%, all the four runs are far from the expected value of 38.6% modifications (recall that 61.4% of the test items are perfect translations). Systems' precision is also lower compared to the English-German task. The average proportion of improved sentences is 23.58%, while the deteriorated ones are on average 55.28%, thus confirming the higher difficulty of the English-Russian evaluation setting.

Overall, the analysis confirms that correcting high-quality translations still remains a hard task, especially when dealing with higher-quality English-Russian outputs. On one side, systems' low precision is an evident limitation. On the other side, one possible explanation is that the margins of improvement to the input sentences are reduced to types of errors (e.g. lexical choice) on which APE systems are less reliable. The analysis proposed in Section 5.2 aims to explore also this aspect.

5.2 Micro indicators: edit operations

In previous rounds of the APE task, the possible differences in the way systems corrected the test set instances were analyzed by looking at the distribution of the edit operations done by each system (insertions, deletions, substitutions and shifts). Such distribution was obtained by computing the TER between the original MT output and the output of each system taken as reference (only for the primary submissions). This analysis has been performed also this year but it turned out to be scarcely informative, as shown in Figure 3.



(a)

(b)

Figure 3: System behaviour (primary submissions) for **English-German** (a) and **English-Russian** (b) – TER(MT, APE)

For both the subtasks, the differences in system's behaviour are indeed barely visible, mainly due to the fact that, in most of the cases, the submitted neural APE models implement similar solutions (multi-source, Transformer-based models trained with the same in-domain and artificial corpora). All the submitted runs are characterized by a large number of substitutions (on average, 53.6% for English-German and 59.7% for English-Russian), followed by the deletions (22.6% for English-German and 26.4% for English-Russian), the insertions (respectively 16.3% and 9.4%) and finally the shifts (7.4% and 4.5%). These results are in line with previous findings. Also in 2018, for instance, the high fluency of neural translations induced the trained models to perform few reordering operations leaving lexical choice as a main direction of improvement, as suggested by the larger amount of substitutions performed by all the systems.

6 Human evaluation

In order to complement the automatic evaluation of APE submissions, a manual evaluation of the primary systems submitted (seven for English-German, five for English-Russian) was conducted. Similarly to the manual evaluation carried out for last year APE shared task, it was based on the direct assessment (DA) approach (Graham et al., 2013; Graham et al., 2017). In this Section, we present the evaluation procedure as well as the results obtained.

6.1 Evaluation procedure

The manual evaluation carried out this year involved 32 native German speakers with full professional proficiency in English. All annotators were paid consultants, sourced by a linguistic service provider company. Each evaluator had experience with the evaluation task through previous work using the same evaluation platform in order to be familiar with the user interface and its functionalities. A screenshot of the evaluation interface is presented in Figure 4.

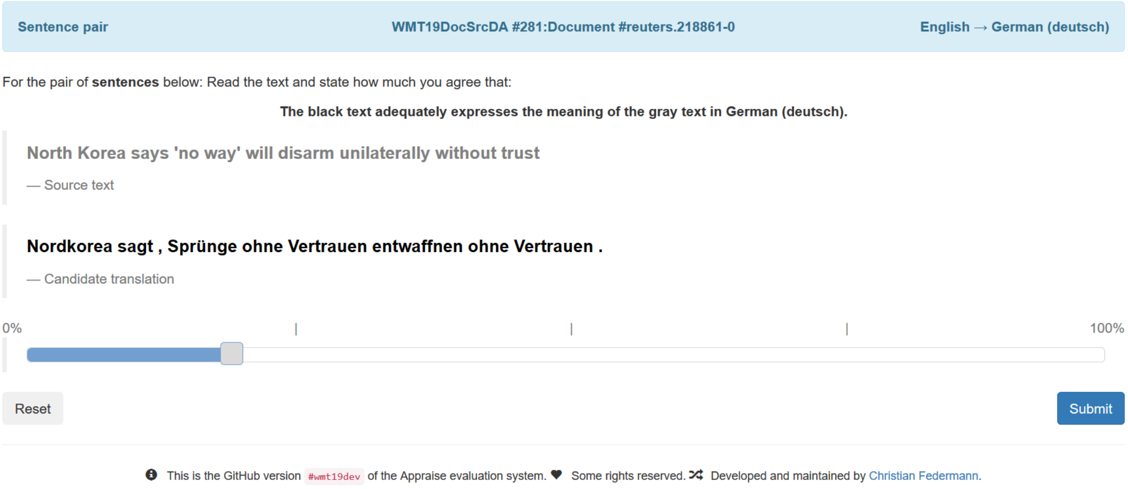
We measure post-editing quality using *source-based direct assessment* (src-DA), as implemented in Appraise (Federmann, 2012). Scores are collected as $x \in [0, 100]$, focusing on adequacy (and not fluency, which previous WMT evaluation campaigns have found to be highly correlated with adequacy direct assessment results).

The original DA approach (Graham et al., 2013; Graham et al., 2014) is reference-based and, thus, needs to be adapted for use in our paraphrase assessment and translation scoring scenarios. Of course, this makes translation evaluation more difficult, as we require bilingual annotators. Src-DA has previously been used, e.g., in (Cettolo et al., 2017; Bojar et al., 2018).

Direct assessment initializes mental context for annotators by asking a priming question. The user interface shows two sentences:

- the source (src-DA, reference otherwise); and
- the candidate output.

Annotators read the priming question and both sentences and then assign a score $x \in [0, 100]$ to the candidate shown. The interpretation of this score considers the context defined by the priming question, effectively allowing us to use the same annotation method to collect assessments wrt. the different dimensions of quality as defined above. Our priming questions are shown in Table 8.



Sentence pair WMT19DocSrcDA #281:Document #reuters.218861-0 English → German (deutsch)

For the pair of **sentences** below: Read the text and state how much you agree that:

The black text adequately expresses the meaning of the gray text in German (deutsch).

North Korea says 'no way' will disarm unilaterally without trust

— Source text

Nordkorea sagt , Sprünge ohne Vertrauen entwaffnen ohne Vertrauen .

— Candidate translation

This is the GitHub version `#wmt19dev` of the Appraise evaluation system. Some rights reserved. Developed and maintained by Christian Federmann.

Figure 4: Screenshot of the direct assessment user interface.

Eval mode	Priming question used
Post-editing adequacy	How accurately does the above candidate text convey the original semantics of the source text? Slider ranges from *Not at all* (left) to *Perfectly* (right).

Table 8: Priming question used for human evaluation of post-editing adequacy.

For adequacy, we ask annotators to assess semantic similarity between source and candidate text, labeled as “source text” and “candidate translation”, respectively. The annotation interface implements a slider widget to encode perceived similarity as a value $x \in [0, 100]$. Note that the exact value is hidden from the human, and can only be guessed based on the positioning of the slider. Candidates are displayed in random order, preventing bias.

For our human evaluation campaign, we also include human post-editing output (`test.tok.pe`) and unedited, neural machine translation output (`test.tok.nmt`). We expect human post-editing to be of higher quality than output from automatic post-editing submissions, which in turn should outperform unedited, neural machine translation output.

6.2 Human Evaluation results

English-German subtask. Score convergence over time for English-German assessments is presented in Figure 5. This figure tracks average system adequacy (as measured by Src-DA) over time, as assessments come in from human annotators. Note that we use the so-called *alternate HIT layout* as named in the WMT18 findings paper, using an 88:12 split between actual assessments and those reserved for quality control. All annotators have proven reliable, passing qualification tests.

The results of Src-DA for the English-German subtask are presented in Table 9. Our main findings are as follows:

- Human post-editing outperforms all auotmatic post-editing systems, the quality difference is significant;
- UNBABEL achieves best APE performance;
- USAAR_DFKI comes in second;
- POSTECH comes in third;
- All but one APE systems outperform unedited NMT output;
- Difference to the remaining APE system is not statistically significant.

Human evaluation does only result in very coarse result cluster. Thus, in order to order submissions by their respective post-editing quality, as perceived by human annotators, we also present *win-based results* in Table 10.

English-Russian subtask. For 2019, we did not run any human evaluation for the English-Russian subtask, due to lack of participation. Instead, we focused annotation efforts on English-German.

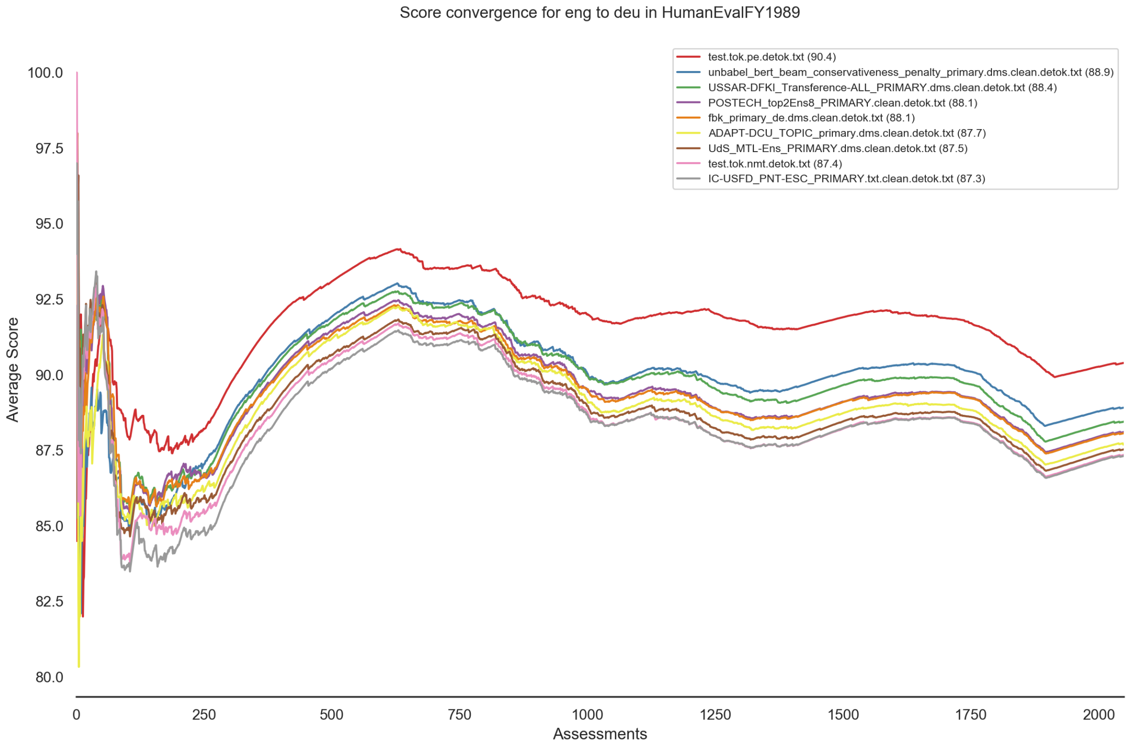


Figure 5: Score convergence over time for English-German assessments.

#	Systems	Ave %	Ave z
1	Human post-edit	90.39	0.154
2	UNBABEL	88.87	0.056
	USAAR_DFKI	88.45	0.027
	POSTECH	88.11	-0.006
	FBK	88.05	-0.014
	ADAPT_DCU	87.70	-0.037
	UDS	87.54	-0.043
	NMT output	87.35	-0.054
	IC_USFD	87.31	-0.059

Table 9: DA Human evaluation results for the **English-German subtask** in terms of average raw DA (Ave %) and average standardized scores (Ave z). Dashed lines between systems indicate clusters according to Wilcoxon signed-rank test at p-level $p \leq 0.05$.

#	Systems	Wins	Ave %	Ave z
1	Human post-edit	8	90.39	0.154
2	UNBABEL	4	88.87	0.056
3	USAAR_DFKI	3	88.45	0.027
4	POSTECH	1	88.11	-0.006
5	FBK	0	88.05	-0.014
	ADAPT_DCU	0	87.70	-0.037
	UDS	0	87.54	-0.043
	NMT output	0	87.35	-0.054
	IC_USFD	0	87.31	-0.059

Table 10: DA Human evaluation results for the **English-German subtask** in terms of average raw DA (Ave %) and average standardized scores (Ave z). Dashed lines between systems indicate clusters according to number of wins.

7 Conclusion

We presented the results from the fifth shared task on Automatic Post-Editing. This year, we proposed two subtasks in which the neural MT output to be corrected was respectively generated by an English-German system and by an English-Russian system. Both the subtasks dealt with data drawn from the information technology domain. Seven teams participated in the English-German task, with a total of 18 submitted runs, while two teams participated in the English-Russian task submitting two runs each. Except in one case (a contrastive run produced with a phrase-based system), the submissions are based on neural approaches, which confirm to be the state-of-the-art in APE. Most of them rely on multi-source models built upon the Transformer and trained by taking advantage of the synthetic corpora released as additional training material.

For the English-German subtask the evaluation was carried out on the same test set used last year, whose human post-edits were not released for the sake of future comparisons. The results on these data, indicate further technology improvements with respect to the 2018 round. This is shown by: *i)* the top result (-0.78 TER and +1.23 BLEU points over the baseline), which is significantly better than last year (-0.38 TER and +0.8 BLEU), and *ii)* the fact that four teams achieved higher results than last year's winning system.

The newly proposed English-Russian subtask proved to be more challenging. None of the submitted runs was able to beat the baseline, whose high TER (16.16) and BLEU (76.2) indicate a very high quality of the initial translations. This is also confirmed by the very skewed TER distribution of the test set items. With more than 60.0% of the translations with TER=0 (the highest value across all the APE datasets released so far), the chance of damaging a perfect MT output is extremely high. Despite the high repetition rate of the English-Russian data (also in this case, the highest across all datasets), the difficulty of handling such a high level of quality contributes to explain the lower results achieved by the two participating teams.

Overall, also this year the main open problem remains to mitigate systems' tendency towards over-correction. In light of the steady progress of NMT technology, handling increasingly better translations calls for conservative and precise solutions able to avoid the unnecessary modification of correct MT output.

Acknowledgments

We would like to thank Apple and Microsoft for their support and sponsorship in organizing the 2019 APE shared task. We are also particularly grateful to the anonymous Appraise annotators that contributed their human intelligence to this activity.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint arXiv:1409.0473*.

David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:2003.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany, August. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 Conference on Machine Translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark, September. Association for Computational Linguistics.

Ondej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (wmt18). In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 272–307, Belgium, Brussels, October. Association for Computational Linguistics.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the iwslt 2017 evaluation campaign. In *Proc. of IWSLT*, Tokyo, Japan.

Rajen Chatterjee, Marion Weller, Matteo Negri, and Marco Turchi. 2015. 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.

Rajen Chatterjee, Matteo Negri, Raphael Rubino, and Marco Turchi. 2018a. Findings of the WMT 2018 shared task on automatic post-editing. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 710–725, Belgium, Brussels, October. Association for Computational Linguistics.

Rajen Chatterjee, Matteo Negri, Marco Turchi, Frédéric Blain, and Lucia Specia. 2018b. Combining quality estimation and automatic post-editing to enhance machine translation output. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 26–38, Boston, MA, March. Association for Machine Translation in the Americas.

Yong Cheng, Zhaopeng Tu, Fandong Meng, Junjie Zhai, and Yang Liu. 2018. Towards robust neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1766, Melbourne, Australia, July. Association for Computational Linguistics.

Gonçalo Correia and André Martins. 2019. A simple and effective approach to automatic post-editing with transfer learning. In *To appear at Proceedings of the 57th annual meeting on association for computational linguistics*. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

M Amin Farajian, Marco Turchi, Matteo Negri, and Marcello Federico. 2017. Multi-domain neural machine translation through unsupervised adaptation. In *Proceedings of the Second Conference on Machine Translation*, pages 127–137.

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

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

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

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Log-linear Combinations of Monolingual and Bilingual Neural Machine Translation Models for Automatic Post-Editing. In *Proceedings of the First Conference on Machine Translation*, pages 751–758, Berlin, Germany, August.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2018. Microsoft and University of Edinburgh at WMT2018: Dual-Source Transformer for Automatic Post-Editing. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium, October.

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

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic.

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

WonKee Lee, Jaehun Shin, and Jong-Hyeok Lee. 2019. Transformer-based Automatic Post-Editing Model with Joint Encoder and Multi-source Attention of Decoder. In *Proceedings of the Fourth Conference on Machine Translation (WMT19)*, Florence, Italy, August.

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

António V. Lopes, M. Amin Farajian, Gonçalo M. Correia, Jonay Trénous, and André F. T. Martins. 2019. Unbabel's Submission to the WMT2019 APE Shared Task: BERT-based Encoder-Decoder for Automatic Post-Editing. In *Proceedings of the Fourth Conference on Machine Translation (WMT19)*, Florence, Italy, August.

Matteo Negri, Marco Turchi, Rajen Chatterjee, and Nicola Bertoldi. 2018. eSCAPE: a Large-scale Synthetic Corpus for Automatic Post-Editing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018.

Santanu Pal, Nico Herbig, Antonio Krüger, and Josef van Genabith. 2019. USAAR-DFKI – The Transference Architecture for English–German Automatic Post-Editing. In *Proceedings of the Fourth Conference on Machine Translation (WMT19)*, Florence, Italy, August.

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

Ofir Press and Lior Wolf. 2017. Using the Output Embedding to Improve Language Models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163, Valencia, Spain, April. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 1715–1725, Berlin, Germany, August.

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

Dimitar Shterionov, Joachim Wagner, and do Carmo Félix. 2019. APE through neural and statistical MT with augmented data. ADAPT/DCU submission to the WMT 2019 APE Shared task. In *Proceedings of the Fourth Conference on Machine Translation (WMT19)*, Florence, Italy, August.

Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007. Statistical Phrase-Based Post-Editing. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT)*, pages 508–515, Rochester, New York.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, USA.

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

Amirhossein Tebbifakhr, Ruchit Agrawal, Rajen Chatterjee, Matteo Negri, and Marco Turchi. 2018. Multi-source Transformer with Combined Losses for Automatic Post-Editing. In *Proceedings of the Third Conference on Machine Translation*, Brussels, Belgium, October.

Amirhossein Tebbifakhr, Matteo Negri, and Marco Turchi. 2019. Effort-Aware Neural Automatic Post-Editing. In *Proceedings of the Fourth Conference on Machine Translation (WMT19)*, Florence, Italy, August.

Marco Turchi, Matteo Negri, and Marcello Federico. 2013. Coping with the subjectivity of human judgements in MT quality estimation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 240–251, Sofia, Bulgaria, August. Association for Computational Linguistics.

Marco Turchi, Matteo Negri, and Marcello Federico. 2014. Data-driven annotation of binary MT quality estimation corpora based on human post-editions. *Machine Translation*, 28(3):281–308.

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

Hongfei Xu and Qiuhui Liu. 2019. Neutron: An Implementation of the Transformer Translation Model and its Variants. *arXiv preprint arXiv:1903.07402*, March.

Hongfei Xu, Qiuhui Liu, and Josef van Genabith. 2019. UdS Submission for the WMT 19 Automatic Post-Editing Task. In *Proceedings of the Fourth Conference on Machine Translation (WMT19)*, Florence, Italy, August.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium, October-November. Association for Computational Linguistics.

Barret Zoph and Kevin Knight. 2016. Multi-source neural translation. *arXiv preprint arXiv:1601.00710.*