Findings of the WMT 2020 Shared Task on Automatic Post-Editing

Rajen Chatterjee(1), **Markus Freitag**(2), **Matteo Negri**(3), **Marco Turchi**(3)
(1) Apple Inc., Cupertino, CA, USA
(2) Google Research, Mountain View, CA, USA
(3) Fondazione Bruno Kessler, Trento, Italy

Abstract

We present the results of the 6^{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 existing human corrections of different sentences. This year, the challenge consisted of fixing the errors present in English Wikipedia pages translated into German and Chinese by state-of-the-art, not domain-adapted neural MT (NMT) systems unknown to participants. Six teams participated in the English-German task, submitting a total of 11 runs. Two teams participated in the English-Chinese task submitting 2 runs each. Due to *i)* the different source/domain of data compared to the past (Wikipedia vs Information Technology), *ii)* the different quality of the initial translations to be corrected and *iii)* the introduction of a new language pair (English-Chinese), this year's results are not directly comparable with last year's round. However, on both language directions, participants' submissions show considerable improvements over the baseline results. On English-German, the top-ranked system improves over the baseline by -11.35 TER and +16.68 BLEU points, while on English-Chinese the improvements are respectively up to -12.13 TER and +14.57 BLEU points. Overall, coherent gains are also highlighted by the outcomes of human evaluation, which confirms the effectiveness of APE to improve MT quality, especially in the new generic domain selected for this year's round.

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

This year, the task focused on two language pairs: English-German and English-Chinese. The former has been part of the APE evaluation campaigns since 2016 (Bojar et al., 2016), while the latter represents a new entry. A second difference with respect to previous rounds is that, for both language pairs, the source/domain of the data changed from Information Technology (IT) to Wikipedia articles. The third major novelty factor consists in the type of MT systems used to generate the translations to be corrected. Although for the third year in a row the task focused on translations produced by neural MT (NMT) systems, this year these models were not adapted to the target domain.

These radical changes have advantages and disadvantages. On one side, moving away from the

Proceedings of the 5th Conference on Machine Translation (WMT), pages 646–659
Online, November 19–20, 2020. ©2020 Association for Computational Linguistics

"narrow" IT domain allowed to test APE technology on the challenging scenario represented by the generic domain of Wikipedia articles. Indeed, as shown in the previous rounds of the task (Chatterjee et al., 2018a, 2019), the high level of repetitiveness of IT data makes this domain easier to model compared to a generic and less repetitive domain, both for MT and APE technology. Moreover, fixing the output of generic NMT models that are not domain-adapted allowed to test APE on lower-quality initial data and verify its potential as a downstream domain adaptation component. On the other side, the disadvantage of changing domain is the reduced possibility to compare results and measure progress across years. Specifically, the lower quality of the original sentences to be corrected (and, in turn, the larger room for improvement left to APE) make the participants' results and the overall technology advancements difficult to analyze in the light of previous rounds.

Six teams participated in the English-German task, submitting eleven runs in total. Two teams participated in the English-Chinese task, submitting two runs each. Similar to last year, all teams developed their systems based on neural technology, which confirms to be the state-of-the-art approach to APE. In most of the cases (see Section 3), participants experimented with the Transformer architecture (Vaswani et al., 2017), either directly or by adapting it to the task. As in previous rounds, their systems exploit information both from the MT output to be corrected and from the corresponding source sentence. This was done either by concatenating the two, as in last year's winning system (Lopes et al., 2019), or by means of multi-source solutions (Zoph and Knight, 2016) successfully explored in the past (Libovický et al., 2016; Chatterjee et al., 2017). Following the recent trends in other NLP areas, the integration of pre-trained BERT-like language models was also considered. Model ensembling and the integration of word/sentence-level quality estimation techniques geared to APE (similar to (Chatterjee et al., 2018b)) were also explored. Finally, also this year participants took advantage of data augmentation techniques, either by creating their own eSCAPE-like corpora (Negri et al., 2018), or by generating synthetic data by adding artificial noise to simulate post-editing errors, or by exploiting external MT candidates as a source of auxiliary information to be concatenated to the input.

The overall evaluation results show significant improvements over the baseline on both the language directions. On **English-German**, where the "*do-nothing*" baseline (see Section 2.3) was 31.56 TER (Snover et al., 2006) and 50.21 BLEU (Papineni et al., 2002), the top-ranked system (20.21 TER, 66.89 BLEU) shows an impressive -11.35 TER reduction, which corresponds to a +16.68 gain in terms of BLEU score. Considering all the submissions, the average gain is -4.89 TER and +6.5 BLEU points, with only one system performing slightly worse than the baseline. Different from last year, where the differences between the top four submissions were not statistically significant, this year we have a clear winner, whose best submission is 6.78 TER points (and 11.12 BLEU points) above the second ranked team. Nevertheless, though possibly favoured by the relatively low baseline results (+14.72 TER and -24.52 BLEU compared to last year), the globally good performance of the participants is a good indicator of overall progress.

The newly proposed **English-Chinese** task is no exception. Here, both participating teams were able to outperform the baseline (59.49 TER and 23.12 BLEU) by a significant margin. The largest gains are up to -12.13 TER and +14.57 BLEU points and, on average for the four submitted runs, they are -8.15 TER and +10.1 BLEU points.

The good results observed with automatic metrics on both the language pairs are confirmed by the human evaluation outcomes. On English-German, for the first time, the top-ranked primary submission is not significantly worse compared to the human post-edited output (suggesting that automatic corrections are indistinguishable from the human ones[1]). All the other systems except one, moreover, are significantly better than the baseline. This also happens for the two primary submissions to the English-Chinese subtask which, however, are both significantly worse than human post-edits.

All in all, the improvements observed on both the language pairs can be most likely ascribed to the lower quality of the initial translations to be corrected. On English-German, the baseline (31.56 TER, 50.21 BLEU) was indeed much lower

[1] A number of factors (related to this year's data and the overall evaluation setting) may have determined this quite surprising finding. Far from claiming to have reached the "human parity" on the APE task, we leave this aspect to future deeper analyses.

than in the past, when the MT systems used were always domain-adapted and hence more competitive. Last year, for instance, the baseline was 16.84 TER (74.73 BLEU), while in none of the previous rounds focusing on this language pair participants had to confront with TER above 25.0 and BLEU below 62.0. On English-Chinese, the baseline was even lower (59.49 TER, 23.12 BLEU), with the lowest scores across all the past six editions of the APE task. On one side, the large gains observed are in line with (and indirectly confirm) previous observations (Bojar et al., 2017; Chatterjee et al., 2018a, 2019) about the difficulty to improve high-quality MT output. Conversely, as we can observe this year, translations of lower quality (like those coming from generic, not domain-adapted models) leave to APE technology a large margin for improvement. On the other side, the observed global gains in both settings motivate further research on APE as a tool for downstream MT adaptation in black-box conditions.

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

For both English-German and English-Chinese, the initial data were selected from English Wikipedia articles and then automatically translated in the two target languages. Although the original English Wikipedia pages were the same, the source sentences eventually used to build the datasets for the two language pairs are different as they were randomly selected.

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/Chinese translation of the source, which was produced by a generic, black-box system unknown to participants. For both the languages, translations were obtained from neural MT systems.[2]
- The human post-edit (PE) is a tokenized 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 *training*, *development* and *test* sets respectively contain 7,000, 1,000 and 1,000 triplets. Participants were also provided with two additional training resources, which were widely used in the previous rounds. One 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.

Also for the **English-Chinese** subtask, the *training*, *development* and *test* sets respectively contain 7,000, 1,000 and 1,000 triplets. For this language pair, however, no additional training resources were provided.

2.1.1 Complexity indicators: repetition rate

Table 1 provides a view of the data from a task difficulty standpoint. For each dataset released in the six rounds of the APE task, it shows the repetition rate of SRC, TGT and PE elements, the TER (Snover et al., 2006) and the BLEU score (Papineni et al., 2002) of the TGT elements (i.e. the original target translations), as well as the TER difference (δ TER) between the top-ranked submission and the task baseline.

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, 2017; Chatterjee et al., 2018a),

[2] Both the NMT systems are based on the standard Transformer architecture (Vaswani et al., 2017) and follow the implementation details described in (Ott et al., 2018). They were trained on publicly available MT datasets including Paracrawl (Esplà et al., 2019) and Europarl (Koehn, 2005), summing up to 23.7M parallel sentences for English-German and 22.6M for English-Chinese.

	2015	2016	2017	2017	2018	2018	2019	2019	2020	2020
Language	En-Es	En-De	En-De	De-En	En-De	En-De	En-De	En-Ru	En-De	En-Zh
Domain	News	IT	IT	Medical	IT	IT	IT	IT	Wiki	Wiki
MT type	PBSMT	PBSMT	PBSMT	PBSMT	PBSMT	NMT	NMT	NMT	NMT	NMT
Rep. Rate SRC	2.905	6.616	7.216	5.225	7.139	7.111	7.111	18.25	0.653	0.81
Rep. Rate TGT	3.312	8.845	9.531	6.841	9.471	9.441	9.441	14.78	0.823	1.27
Rep. Rate PE	3.085	8.245	8.946	6.293	8.934	8.941	8.941	13.24	0.656	1.2
Baseline TER	23.84	24.76	24.48	15.55	24.24	16.84	16.84	16.16	31.56	59.49
Baseline BLEU	n/a	62.11	62.49	79.54	62.99	74.73	74.73	76.20	50.21	23.12
δ TER	+0.31	-3.24	-4,88	-0,26	-6.24	-0.38	-0.78	+0.43	-11.35	-12.13

Table 1: 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). The last row (δ TER) indicates, for each evaluation round, the difference in TER between the baseline (i.e. the "*do-nothing*" system) and the top-ranked submission.

suggest a higher chance of learning from the training set correction patterns that are applicable also to the test set.

Over the years, the relation between systems' performance and the repetition rate observed in the data has been analysed in the light of the different values reported in Table 1. Some rounds of the task suggested the hypothesis that large differences in repetitiveness across the datasets give a possible explanation for the variable gains over the baseline achieved by participants. Indeed, in some cases (e.g. in the APE15 task and in the APE17 German-English subtask), low repetition rates seemed to motivate generally low systems' results, while in others (e.g. APE17 English-German subtask) also the opposite was true, with large gains over the baseline associated to high repetition rates. However, the outcomes of other rounds of the task do not support this intuition. In the 2018 round, despite the relatively high repetition rate values observed in the data, evaluation results shown that the influence of data repetitiveness on final APE performance is marginal. The same happened in 2019 (Chatterjee et al., 2019), when the highest repetition rates ever measured in the APE data (English-Russian subtask) were not enough to develop systems able to improve over the baseline.

As discussed in Section 4, this year we are in the opposite situation. On both English-German and English-Chinese, the lowest repetition rates ever measured in the APE data did not prevent participants from achieving considerable gains over the baseline. This confirms that, as hypothesized last year, systems' improvements over the baseline are either scarcely correlated to text repetitiveness or more influenced by other task difficulty 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 (Baseline TER/BLEU rows in Table 1) using the human post-edits as reference.

As discussed in (Bojar et al., 2017; Chatterjee et al., 2018a, 2019), 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. Similarly, in 2019 the very high translation

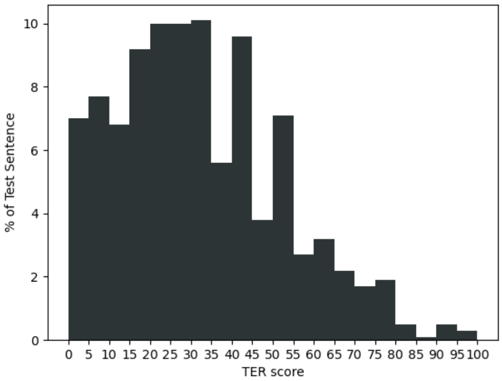


Figure 1: TER distribution in the **English-German** test set

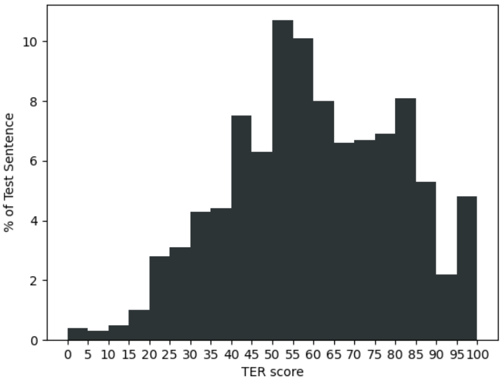


Figure 2: TER distribution in the **English-Chinese** test set

quality featured by strong, domain-adapted neural models made the task rather difficult. On English-German, where the baseline system was again very competitive (16.84 TER, 74.73 BLEU), the largest TER reduction was indeed of 0.78 points (corresponding to a BLEU increase of 1.23). On English-Russian, where the initial MT quality was even higher,[3] (16.16 TER, 76.2 BLEU), the baseline remained unbeaten.

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. Indeed, with baseline TER and BLEU scores significantly lower than in all the other rounds of the task (31.56 TER and 50.21 BLEU for English-German, 59.49 TER and 23.12 BLEU for English-Chinese), almost all participants managed to obtain very large improvements despite the low repetition rates featured by the data.

2.1.3 Complexity indicators: TER distribution

A third complexity indicator considered in previous rounds of the task is the TER distribution (computed against human references) for the translations present in the test sets. What we observed in the past is that harder tasks were typically characterized by TER distributions particularly skewed towards low values. For instance, in 2019 around 50% of the English-German and 63.5% of the English-Russian test items had a TER between 0 and 10, the latter subtask being considerably more difficult than the former (recall that, on English-Russian, none of the participants was able to beat the baseline). Indeed, the higher the proportion of (near-)perfect test instances (i.e. items with 0<TER<10, which hence require few edits or no corrections at all), the higher the probability that APE systems will perform unnecessary corrections that will be penalized by automatic evaluation metrics.

On the contrary, less skewed distributions can be expected to be easier to handle as they give to automatic systems a larger room for improvement. In the lack of more focused analyses on this aspect, we can hypothesize that, in ideal conditions from the APE standpoint, the peak of the distribution would be observed for "post-editable" translations containing enough errors that leave some margin for focused corrections, but not too many errors to be so unintelligible to require a whole re-translation from scratch.[4]

As shown in Figures 1 and 2, the TER distributions in the two test sets released this year is quite different from previous rounds and actually reflects a more balanced situation. For English-German, about 55% of the samples falls in the 15-45 TER interval, with no more $\sim$ 7% of the items being perfect (i.e. TER=0). For English-Chinese, for which the overall MT quality is significantly lower (as shown by the worse baseline results reported in Table 1), the vast majority of the samples falls in the 40-85 interval, with less than 1% of the

[3] Note that the higher quality of the initial translations added up to the higher difficulty of dealing with a morphologically-rich target language like Russian. The two aspects are clearly tightly connected and disentangling them would require further analysis. Nonetheless, regarding the correlation between MT quality and final results, also this subtask was not an exception compared to the other settings summarized in Table 1.

[4] For instance, based on the empirical findings reported in (Turchi et al., 2013, 2014), TER=0.4 is the threshold that, for human post-editors, separates the "post-editable" translations from those that require complete rewriting from scratch.

ID	Participating team
MinD	Alibaba Group, Hangzhou, China (Wang et al., 2020)
BeringLab	Bering Lab, Republic of Korea (Lee, 2020)
HW-TSC	Huawei Translation Services Center & East China Normal University, China (Yang et al., 2020)
KAIST	Korea Advanced Institute of Science & Technology, Republic of Korea
POSTECH	Pohang University of Science and Technology, Republic of Korea (Lee et al., 2020b)
POSTECH-ETRI	Pohang University & Electronics and Telecomm. Res. Inst., Republic of Korea (Lee et al., 2020a)

Table 2: Participants in the WMT20 Automatic Post-Editing task.

items being perfect.

In the light of previous years' observations, both the subtasks hence seem to be easier to handle. As discussed in Section 4, also this year's evaluation results confirm the strict correlation between the quality of the initial translations, the distribution of TER scores across the test items, 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[5] software: lower average TER scores correspond to higher ranks. BLEU was computed using the multi-bleu.perl package[6] available in MOSES. The evaluation results computed in terms of automatic metrics are presented and discussed in Section 4).

Manual evaluation was conducted via source-based direct human assessment (Graham et al., 2013; Cettolo et al., 2017; Bojar et al., 2018). 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 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 1, are also reported in Tables 3 and 4 for comparison with participants' submissions.[7]

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

Six teams submitted a total of 11 runs for the English-German subtask. Two of them participated also in the English-Chinese subtask by submitting 2 runs each. Participants are listed in Table 2, and a short description of their systems is provided in the following.

Alibaba Group (*MinD*). Alibaba participated only in the English-German subtask. Their submission introduces a cross-lingual Bert-like conditional model with a "memory-encoder", which can capture the semantic information of machine translations conditional on the source sentences (Fan et al., 2019). The system consists of three parts, namely: *i)* a general Transformer encoder to encode the source sentences, *ii)* a Transformer decoder without future mask adapted to a memory-encoder to encode machine translations with cross attention to the source encoder, and *iii)* a multi-source Transformer decoder to generate the automatic post-editing results with cross attentions to both the encoders. In addition, data augmentation, corpus filtering and imitation learning strategies are exploited to overcome the scarcity of real APE data and to further improve model's performance, together with model ensembling and conservativeness penalty strategies inspired by (Lopes et al., 2019).

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

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

[7] 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, 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. Since 2018, we hence opted for considering only one baseline.

Bering Lab (*BerlingLab*). Bering Lab participated only in the English-German subtask. Their system relies on a Transformer architecture, in which the encoder takes in input a concatenation of the source and MT sentences to generate a cross-lingual representation to be passed to the decoder. Additionally, they explored methods to improve APE performance through word-level and sentence-level quality estimation (QE). Based on word-level QE, they mask incorrect or missing words in the PE output. Then, the most probable word for each masked token is predicted using XLM-RoBERTa (Conneau et al., 2020), which is fine-tuned based on the translation language modeling (TLM) objective (Conneau and Lample, 2019). Finally, they propose an output selection mechanism based on sentence-level QE to prevent over-correction. To this aim, they select the sentence with the lowest predicted HTER among the PE outputs and the original MT sentence as the final output. For data augmentation, they use a parallel corpus to train an NMT model and generate artificial triplets, following the ideas from (Negri et al., 2018).

Huawei (*HW-TSC*). Huawei participated both in the English-German and English-Chinese subtasks. Their system basically follows the architecture of last year's winning system (Lopes et al., 2019), where *src* and *mt* sentences are concatenated as input to the encoder, and the decoder is used for decoding the *pe* sentence. However, there are several differences with respect to (Lopes et al., 2019). First, instead of using a pre-trained BERT model, the system relies on a Transformer NMT model (implemented with fairseq (Ott et al., 2019)), pre-trained on the WMT19 news translation corpora. Second, the model integrates bottleneck adapter layers to prevent from over-fitting. Third, external MT candidates (from Google Translate) are exploited as a source of auxiliary information. This results in a longer input sequence composed of (*src*, *mt*, *auxiliary_mt*) triplets. Due to the domain change introduced this year, system's training does not exploit the supplied additional corpora for data augmentation. Finally, the system does not include methods to prevent over-correction, such as the penalty mentioned in (Lopes et al., 2019).

POSTECH (*POSTECH_TERNoise*). This team participated only in the English-German subtask. They mainly focused on increasing the size of the APE data to overcome the scarcity of training samples available. They first introduced a noising module simulating the four types of post-editing errors: insertion, deletion, substitution and shifting. This noising module implants the simulated errors into the target text of the parallel corpora, so to exploit a synthetic MT output during the training phase. The quantity of noise is determined by using the TER distribution of the official training set. They then applied the same generation method proposed in (Negri et al., 2018), so to create a synthetic APE corpus to be used as additional training data. For this data construction process, they used the parallel corpora and the NMT model released for the WMT20 Quality Estimation shared task. As APE model, they chose the sequential model proposed in (Lee et al., 2019), applying some minor modifications to increase the training efficiency. They submitted two ensemble models. Their primary submission (TERNoise-Ops-Ens8) is an ensemble of eight runs. It was obtained by first selecting the top-5 runs having the lowest TER on the development set, for three individual weight initializations. Out of them, they then selected the top-2 runs showing most frequent corrections for each of the four edit operations to form the ensemble. The contrastive submission (TERNoise-nFold-Ens8) is an ensemble of eight runs obtained from models trained/validated in a 4-fold setting on the integration of training data and development data, aiming at the generalization to unseen data. Then, the top-2 runs for each fold were selected to form the ensemble.

POSTECH-ETRI (*POSTECH-ETRI*). This team participated both in the English-German and English-Chinese subtasks. Their models focus on adapting to the APE task XLM (Conneau and Lample, 2019), which can learn joint representations from two languages. Rather than using the open model published on the XLM github page[8] trained on 15 languages, they built new MLM+TLM models that are trained on datasets consisting of only the source and target languages for both language pairs (English-German and English-Chinese). Their model architecture is an extension of Transformer, in which the encoder is initialized with the weights of the pre-trained

[8] https://github.com/facebookresearch/XLM

		TER	BLEU
en-de	HW-TSC_DIRECT_CONTRASTIVE.pe	20.21	66.89
	HW-TSC_CONCAT_PRIMARY.pe	20.52	66.16
	MinD-mem_enc_dec_post-CONTRASTIVE	26.99	55.77
	POSTECH-ETRI_XLM-Top4Ens_CONTRASTIVE	27.02	56.37
	MinD-mem_enc_dec-PRIMARY	27.03	55.58
	POSTECH-ETRI_XLM-Top3Ens_PRIMARY	27.37	55.83
	BeringLab_model1_PRIMARY	27.61	54.71
	BeringLab_model2_CONTRASTIVE	27.96	54.60
	POSTECH_TERNoise-nFold-Ens8_CONTRASTIVE	28.22	54.51
	POSTECH_TERNoise-Ops-Ens8_PRIMARY	28.41	54.22
	Baseline	31.56	50.21
	KAISTxPAPAGO_EMT_PRIMARY	32.00	49.21

Table 3: Results for the WMT20 APE **English-German** – average TER (↓), BLEU score (↑).

		TER	BLEU
en-zh	HW-TSC_CONCAT_PRIMARY.pe	47.36	37.69
	HW-TSC_DIRECT_CONTRASTIVE.pe	48.01	37.32
	POSTECH-ETRI_XLM-Top3Ens_PRIMARY	54.92	28.90
	POSTECH-ETRI_XLM-Top4Ens_CONTRASTIVE	55.08	28.97
	Baseline	59.49	23.12

Table 4: Results for the WMT20 APE **English-Chinese** – average TER (↓), BLEU score (↑).

XLM, receiving the concatenation of the two input sentences. The decoder is also initialized in a similar manner as the encoder, while multi-head attention layers are random-initialized. At the APE training stage, in addition to the WMT20 official dataset, they generated new synthetic triplets, following the same method used to build eSCAPE (Negri et al., 2018). They used the NMT model provided by the WMT20 quality estimation shared task to generate new synthetic APE triplets by translating the parallel corpus provided by the same task. Finally, to generate their final submissions, they built an ensemble of multiple models.

4 Results

Participants' results are shown in Tables 3 (English-German) and 4 (English-Chinese). 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. The two tables also report the BLEU score computed using human post-edits, which represents our secondary evaluation metric.

Similar to last year, also in this round the primary and secondary evaluation metric produce rankings that are only slightly different from each other.[9] In spite of these minor difference, for both both languages we have a clear separation between the two top-ranked submissions (by the same team) and the other submitted runs.

On **English-German**, the best results (20.21 TER, 66.89 BLEU) respectively outperform the baseline by -11.35 TER and +16.68 BLEU points, the second-best scores being lower by less than 1 point for both the metrics. All the other runs but the last are quite close to each other, being concentrated respectively in a 1.42 TER and 1.55 BLEU points interval.

On **English-Chinese**, the best results (47.36 TER, 37.69 BLEU) respectively outperform the baseline by -12.13 TER and +14.57 BLEU points. Also in this case, the second-best run is below the top-ranked one by less than 1 point for both the metrics, while the third and fourth submissions are close to each other (the difference is less than 0.2 points for both metrics).

All in all, these results indicate that:

- Operating with lower-quality output produced by generic (i.e. not domain-adapted) NMT systems run on a broad "domain" like Wikipedia texts (as opposed to the narrow domains of information technology or medical) leaves considerable room for improvement to state-of-the-art APE models. Looking at the baseline scores and the δTER values shown

[9] For English-German, the third and fourth-ranked submissions in terms of TER are switched in terms of BLEU, as well as the fifth and the sixth. For English-Chinese, this happens for the third and fourth-ranked submissions. The correlation between the ranks obtained by the two metrics is however very high, and in both cases above 0.99.

Systems	Modified	Improved	Deteriorated	Prec.
HW-TSC_DIRECT_CONTRASTIVE.pe	905 (90.5%)	625 (69.06%)	177 (19.56%)	0.69
HW-TSC_CONCAT_PRIMARY.pe	908 (90.8%)	618 (68.06%)	183 (20.15%)	0.68
MinD-mem_enc_dec_post-CONTRASTIVE	662 (66.2%)	397 (59.97%)	148 (22.36%)	0.60
POSTECH-ETRI_XLM-Top4Ens_CONTRASTIVE	771 (77.1%)	438 (56.81%)	199 (25.81%)	0.57
MinD-mem_enc_dec-PRIMARY	665 (66.5%)	401 (60.30%)	144 (21.65%)	0.60
POSTECH-ETRI_XLM-Top3Ens_PRIMARY	778 (77.8%)	423 (54.37%)	207 (26.61%)	0.54
BeringLab_model1_PRIMARY	708 (70.8%)	380 (53.67%)	157 (22.18%)	0.54
BeringLab_model2_CONTRASTIVE	421 (42.1%)	279 (66.27%)	72 (17.10%)	0.66
POSTECH_TERNoise-nFold-Ens8_CONTRASTIVE	535 (53.5%)	306 (57.20%)	108 (20.19%)	0.57
POSTECH_TERNoise-Ops-Ens8_PRIMARY	536 (53.6%)	309 (57.65%)	112 (20.90%)	0.58
KAISTxPAPAGO_EMT_PRIMARY	724 (72.4%)	267 (36.88%)	314 (43.37%)	0.37
Average	69.2	58.2	23.6	0.58

Table 5: Number (raw and proportion) of test sentences modified, improved and deteriorated by each run submitted to the APE 2020 **English-German** subtask. The "Prec." column shows systems' precision as the ratio between the number of improved sentences and the total number of modified instances.

Systems	Modified	Improved	Deteriorated	Prec.
HW-TSC_CONCAT_PRIMARY.pe	997 (99.7%)	673 (67.50%)	227 (22.77%)	0.68
HW-TSC_DIRECT_CONTRASTIVE.pe	995 (99.5%)	671 (67.44%)	223 (22.41%)	0.67
POSTECH-ETRI_XLM-Top3Ens_PRIMARY	968 (96.8%)	566 (58.47%)	265 (27.38%)	0.58
POSTECH-ETRI_XLM-Top4Ens_CONTRASTIVE	959 (95.9%)	551 (57.46%)	255 (26.59%)	0.57
Average	97.975	62.72	24.79	0.63

Table 6: Number (raw and proportion) of test sentences modified, improved and deteriorated by each run submitted to the APE 2020 **English-Chinese** subtask. The "Prec." column shows systems' precision as the ratio between the number of improved sentences and the total number of modified instances.

in Table 1, we can observe that the largest improvements over the baseline were obtained this year on the lowest-quality translations.

- Operating with data featuring low repetition rates does not necessarily prevent from obtaining significant MT quality improvements. Looking at the δTER and the repetition rate values shown in Table 1, we can observe that the lowest data repetitiveness observed this year did not prevent from observing, at the same time, the largest gains over the baseline.

- Operating with data featuring variable quality, with a distribution of the instances that is not too peaked towards high-quality translations, sets ideal conditions for APE. Looking at the δTER and the TER distributions shown in Figures 1 and 2, we can observe that the largest improvements over the baseline achieved this year are also related to a quality distribution that is more uniformly spread around central values of the 0-100 TER interval.

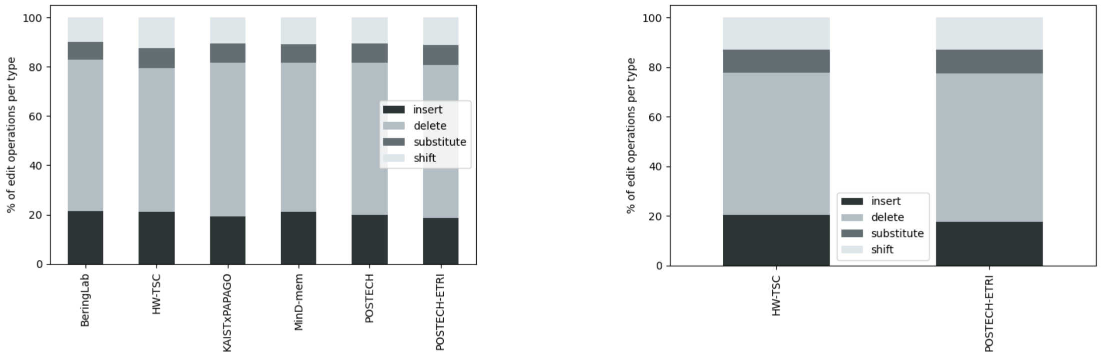
5 System/performance analysis

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 5 and 6 show, for each run submitted to the two subtasks, the number of modified, improved and deteriorated sentences, as well as the overall system's precision (i.e. the proportion of improved sentences out of the total number of modified instances). 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.

As shown in Table 5, on **English-German** the amount of sentences modified by the participants varies from the very high values of the top two submissions (above 90.0%) to the lower scores of the runs placed below them in the ranking (between 42.1% and 77.8%). However, in all the cases the overall number of modified sentences (69.2% on average) is considerably larger than what we observed in the 2019 round (23.53% on average, ranging from 4.01% to 39.1%). This difference can be ascribed to the different nature of the data that, as previously discussed, this year



(a)

(b)

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

featured lower MT quality, combined with a distribution that is less skewed towards low TER values. In particular, while last year about 30.0% of the test instances were to be considered as "perfect", this year the proportion of test instances with 0≤TER≤5 is about 7.0%. In light of this, compared to last year, the participants modified a number of test instances that is much closer to the target percentage of sentences to be modified (about 93.0%, i.e. those having TER>0). As one can expect, besides systems' aggressiveness, final performance highly depends also on their precision in applying corrections. The last column of Table 5 shows systems' precision (Prec.) as the ratio between the number of improved sentences and the total number of modified sentences. As can be seen from the table, the two top-ranked submissions are not only the most aggressive (more than 90% modified sentences) but also the most precise ones (precision above 0.68). Overall, all runs but one have a precision above 0.5, with an average value of 0.58 that is larger than the values observed on the same language (but different evaluation conditions) in 2019 (0.46) and in 2018 (0.34). As a consequence, the percentage of deteriorated sentences out of the total amount of modified test items shows a significant drop with respect to the last two rounds of the task. On average, a quality decrease is observed for 23.6% of the test items (it was 47.85% in 2018 and 35.11% in 2019).

As shown in Table 6, on **English-Chinese** we observe similar trends. The four submitted runs are all characterized by a high percentage of modified sentences (97.97% on average) and a very high precision (0.63 on average). This can be explained by the large room for improvement available to APE on this language pair, due to the low MT baseline (59.49 TER, 23.12 BLEU) and to the small number of "perfect" translations (as shown in Figure 2, less than 0.5% of the test items have a 0≤TER≤5).

5.2 Micro indicators: edit operations

In the previous rounds of the APE task, 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. For both the subtasks, the differences in system's behaviour are indeed barely visible. All the submitted runs are characterized by a large number of deletions (on average, 61.11% for English-German and 58.54% for English-Chinese), followed by the insertions (respectively, 20.17% and 19.01%), the shifts (10.98% and 12.98%) and finally the substitutions (7.74 and 9.48). These distributions differ from what we observed in the past. Especially in the last two rounds of the APE task, the largest proportion of edit operations were indeed substitutions (for English-German neural translations, they were 53.6% in 2019 and 53.5% in 2018). Also this difference can be explained by the lower quality of this year's initial translations. In the previous rounds, the generally high fluency of domain-adapted neural MT systems induced the trained APE models to perform light changes, mainly with isolated word substitutions oriented to improve lexical choice. This year, the change of domain and the use of generic models that were not domain-adapted resulted in more aggressive structural modifications, where lexical changes represent the minority of edit operations.

6 Human evaluation

In order to complement the automatic evaluation of APE submissions, manual evaluation of the primary systems submitted (seven for English-German, three for English-Chinese) was conducted. In this section, we present the evaluation procedure, as well as the results obtained.

6.1 Evaluation procedure

We evaluated the overall quality of the MT and PE output using source-based direct assessment (Graham et al., 2013; Cettolo et al., 2017; Bojar et al., 2018). We used the same instructions that are used in the News Translation track of WMT2020. We hired 25 professional linguists for English-German and 25 professional linguists for English-Chinese. All involved linguists were either native speaker in German or Chinese.

We acquired only a single rating per sentence as we found that professional linguists were more reliable than crowd workers (Toral, 2020). For adequacy, we asked annotators to assess the semantic similarity between the source and a candidate text, labelled as "source text" and "candidate translation", respectively. The annotation interface implements a slider widget to encode perceived similarity as a value between 0 and 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, so to prevent biased assessments.

For our human evaluation campaign, we also include the human post-edits (test.pe) and the unedited, MT output (test.mt). We expect human post-editing to be of higher quality than the output from APE submissions, which in turn should outperform the unedited MT output. We run human evaluation for all primary submissions, the MT output and the human post-edited output.

6.1.1 English→German

Human evaluation results for English-German are summarized in Table 7. The human post-edited output *test.pe* scores best, while the APE output *HW-TSC_CONCAT.pe* is not significantly worse compared to the human post-edited output. Consequently, and rather surprisingly, human and automatic corrections for this language pair seem to be indistinguishable to our evaluators. This interesting finding can be motivated by a number of reasons (the type/quality/quantity of data, the size of the sample, the number of collected judgements) that suggest to avoid exaggerated claims about a reached human parity. Nonetheless, we take it as indicator of a steady progress of APE research. Interestingly, 5 out of 6 APE submissions perform significantly better than the original MT output *test.mt*, demonstrating that APE can be used to improve machine translation output even for high-resource language settings like English-German, as already shown by Freitag et al. (2019). These findings are different from last year's APE task (Chatterjee et al., 2019) where none of the English-German APE submissions was significantly better than the raw MT output.

	Avg	Avg z
test.pe	83.5	0.298
HW-TSC_CONCAT.pe	82.2	0.260
POSTECH-ETRI_XLM-Top3Ens	77.3	0.031
MinD-mem_enc_dec	76.2	-0.008
POSTECH_TERNoise-Ops-Ens8	75.8	-0.037
BeringLab_model1	74.3	-0.098
test.mt	71.5	-0.194
KAISTxPAPAGO_EMT	71.0	-0.252

Table 7: Results for the WMT20 APE **English-German – human evaluation**. Systems ordered by DA score z-score; systems within a cluster are considered tied; lines indicate clusters according to Wilcoxon rank-sum test $p < 0.05$.

	Avg	Avg z
test.pe	86.3	0.363
HW-TSC_CONCAT.pe	77.2	-0.063
POSTECH-ETRI_XLM-Top3Ens	77.0	-0.079
test.mt	74.0	-0.221

Table 8: Results for the WMT20 APE **English-Chinese – human evaluation**. Systems ordered by DA score z-score; systems within a cluster are considered tied; lines indicate clusters according to Wilcoxon rank-sum test $p < 0.05$.

6.1.2 English→Chinese

Human evaluation results for English-Chinese are summarized in Table 8. In this case, the human post-edited output does perform significantly better than the two primary submissions. Similar to the English-German task, both APE submissions perform significantly better than the original MT output *test.mt*. Nevertheless, both submissions perform very similarly, and both submissions can be seen as similar quality.

7 Conclusion

We presented the results from the 6^{th} shared task on Automatic Post-Editing at WMT. This year, we proposed two subtasks in which the MT output to be corrected was respectively generated by English-German and English-Chinese neural systems unknown to the participants. The latter language pair represents a new entry for the task, which previously focused on Spanish (in 2015), German (since 2016) and Russian (in 2019) as target languages. The other major novelty factors are that: *i)* both the subtasks dealt with data drawn from the "generic" domain of Wikipedia articles, and *ii)* the NMT systems used to generate the translations were not domain-adapted. As a consequence, participants had to confront with lower quality translations that left to APE large room for improvement.

Six teams participated in the English-German task, with a total of 11 submitted runs, while two teams participated in the English-Chinese task submitting two runs each. Their results computed with automatic metrics (TER and BLEU) revealed significant gains over the "*do-nothing*" baseline. On English-German, the top-ranked system improved over the baseline by -11.35 TER and +16.68 BLEU points, and the average improvements were the largest ones observed over the years (-4.89 TER, +6.5 BLEU). On English-Chinese the improvements of the top-ranked system are respectively -12.13 TER and +14.57 BLEU points, with average gains of (-8.15 TER and +10.1 BLEU). Our human evaluation confirmed that on both the language pairs, almost all the primary submissions are significantly better than the baseline. On English-German, the improvement is up to the point that the quality of the automatic corrections produced by the top-ranked primary submissions is substantially on par with human corrections.

All in all, these results confirm the effectiveness of APE to improve MT output in black-box conditions, especially when the adaptation of generic systems to a new "domain" is required.

Acknowledgments

We would like to thank Apple and Google Research for their support and sponsorship in organizing the 2020 APE shared task.

References

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. Association for Computational Linguistics.

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

Ondřej Bojar, 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. 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, M. Amin Farajian, Matteo Negri, Marco Turchi, Ankit Srivastava, and Santanu Pal. 2017. Multi-source Neural Automatic Post-Editing: FBK's Participation in the WMT 2017 APE Shared Task. In *Proceedings of the Second Conference on Machine Translation*, pages 630–638. Association for Computational Linguistics.

Rajen Chatterjee, Christian Federmann, Matteo Negri, and Marco Turchi. 2019. Findings of the WMT 2019 shared task on automatic post-editing. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 11–28, Florence, Italy. Association for Computational Linguistics.

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. 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. Association for Machine Translation in the Americas.

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 and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 156–161, Beijing, China. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 7057–7067.

Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland. European Association for Machine Translation.

Kai Fan, Jiayi Wang, Bo Li, Boxing Chen, and N. Ge. 2019. Neural zero-inflated quality estimation model for automatic speech recognition system. *ArXiv*, abs/1910.01289.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. APE at Scale and Its Implications on MT Evaluation Biases. In *Proceedings of the Fourth Conference on Machine Translation*, pages 34–44, Florence, Italy. Association for Computational Linguistics.

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

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.

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

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, pages 79–86, Phuket, Thailand.

Dongjun Lee. 2020. Cross-Lingual Transformers for Neural Automatic Post-Editing. In *Proceedings of the Fifth Conference on Machine Translation*, Online.

Jihyung Lee, WonKee Lee, Jaehun Shin, Baikjin Jung, Young-Kil Kim, and Jong-Hyeok Lee. 2020a. POSTECH-ETRI's Submission to the WMT2020 APE Shared Task: Automatic Post-Editing with Cross-lingual Language Model. In *Proceedings of the Fifth Conference on Machine Translation*, Online.

WonKee Lee, Jaehun Shin, Baikjin Jung, Jihyung Lee, and Jong-Hyeok Lee. 2020b. Noising Scheme for Data Augmentation in Automatic Post-Editing. In *Proceedings of the Fifth Conference on Machine Translation*, Online.

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 (Volume 3: Shared Task Papers, Day 2)*, pages 112–117, Florence, Italy. Association for Computational Linguistics.

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 (Volume 3: Shared Task Papers, Day 2)*, pages 118–123, Florence, Italy. Association for Computational Linguistics.

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. European Language Resources Association (ELRA).

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation (WMT)*.

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.

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.

Antonio Toral. 2020. Reassessing Claims of Human Parity and Super-Human Performance in Machine Translation at WMT 2019. *arXiv preprint arXiv:2005.05738*.

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. 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.

Jiayi Wang, Ke Wang, Kai Fan, Yuqi Zhang, Jun Lu, Xin Ge, Yangbin Shi, and Yu Zhao. 2020. Alibaba's Submission for the WMT 2020 APE Shared Task: Improving Automatic Post-Editing with Pre-trained Conditional Cross-Lingual BERT. In *Proceedings of the Fifth Conference on Machine Translation*, Online.

Hao Yang, Minghan Wang, Daimeng Wei, Hengchao Shang, Jiaxin Guo, Zongyao Li, Lizhi Lei, Ying Qin, Shimin Tao, Shiliang Sun, and Yimeng Chen. 2020. HW-TSC's Participation at WMT 2020 Automatic Post Editing Shared Task. In *Proceedings of the Fifth Conference on Machine Translation*, Online.

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