Findings of the WMT 2022 Shared Task on Automatic Post-Editing

Pushpak Bhattacharyya
IIT Bombay

Rajen Chatterjee
Apple Inc.

Markus Freitag
Google

Diptesh Kanojia
University of Surrey

Matteo Negri
Fondazione Bruno Kessler

Marco Turchi
Zoom Video Communications

Abstract

We present the results from the 8^{th} round of the WMT shared task on MT Automatic Post-Editing, which consists in automatically correcting the output of a "black-box" machine translation system by learning from human corrections. This year, the task focused on a new language pair (English→Marathi) and on data coming from multiple domains (healthcare, tourism, and general/news). Although according to several indicators this round was of medium-high difficulty compared to the past, the best submission from the three participating teams managed to significantly improve (with an error reduction of 3.49 TER points) the original translations produced by a generic neural MT system.

1 Introduction

This paper presents the results of the 8^{th} round of the WMT task on MT Automatic Post-Editing (APE). The task consists in automatically correcting the output of a "black-box" machine translation system by learning from human-revised machine-translated output supplied as training material. The overall task formulation (see Section 2) remained the same as in all previous rounds, where the challenge consisted in fixing the errors present in English documents automatically translated by state-of-the-art, not domain-adapted neural MT (NMT) systems unknown to participants. However, two main factors of novelty characterized the APE 2022 evaluation setting:

- **Language Pair:** This year, we focus on English→Marathi. Marathi is an Indo-Aryan language predominantly spoken by Marathi people in the Indian state of Maharashtra (see Section 3).
- **Data Domain:** Instead of covering one single domain as in previous rounds (either news, medical, or information technology of Wikipedia documents), training/dev/test data were selected from a mix of domains, namely: healthcare, tourism, and general/news.

This year, we had three teams submitting a total of five systems for final evaluation (see Section 5). While the difficulty (Section 4) of this round falls in a medium-high range attested by relatively high baseline results on the test data (20.28 TER / 67.55 BLEU), final results indicate the overall good quality of the submitted runs. Two teams were indeed able to significantly improve over the baseline in terms of the official automatic evaluation metrics (Section 6). In particular, according to the primary metric (*i.e.,* the TER score computed between automatic and human post-edits), the top-ranked system (16.79 TER / 72.92 BLEU) achieved an error reduction of 3.49 TER points. Also, this year, the standard automatic evaluation was complemented by a human evaluation based on direct assessment. However, some problems in the procedure[1] were later discovered, which make it unreliable to draw insights except for the confirmation that two of the three submitted systems were able to improve over the baseline significantly. Specifically, both of them achieved a mean direct assessment score that drastically reduces the gap between the baseline and human post-editing quality. However, due to the mentioned problems in the human evaluation procedure, further details about it will not be included in the discussion below.

Although the different language/domain testing conditions prevent from drawing precise conclusions about the progress of APE technology with respect to last year, the overall positive results confirm its viability for downstream improvements of "black-box" MT systems whose inner workings are not accessible.

[1] Basically, due to an error in assigning the direct assessment tasks, the scores collected can be used to compare systems to the baseline but cannot be used to compare them to each other.

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 109–117

2 Task Description

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.

This 8^{th} round of the WMT APE shared task kept the same overall evaluation setting of the previous seven rounds. Specifically, the participating systems had to automatically correct the output of an unknown "black-box" MT system (a generic NMT system not adapted to the target domain) by learning from training data containing human revisions of translations produced by the same system. The selected language pair and the data domain, however, were totally new to the task. Different from previous rounds covering more language pairs (or directions), this year focused only on English-Marathi, presenting participants with the traditional source language and, for the third time in a row, an Eastern language as the target. Moreover, while the training, development and test data released in previous rounds were always drawn from a single domain, this year, they covered three domains: healthcare, tourism, and general/news.

3 Data, Metrics, Baseline

3.1 Data

In this round of the APE task, we introduce a new language pair - English-Marathi. Marathi is one of the most spoken Indian languages, with approximately 83 million native speakers and 16 million speakers as a second/third language[2]. Marathi is a known agglutinative language and presents various challenges to machine translation when compared to its other Indian counterparts (Khatri et al., 2021; Banerjee et al., 2021). Moreover, the English-Marathi language pair is considered a low-resource language pair compared to English-Hindi/Bengali/Malayalam (Ramesh et al., 2022) despite having more native speakers around the world. An automatic post-editing approach which helps correct the issues posed by NMT systems is crucial for a low-resource language such as Marathi.

As in all previous rounds, participants were provided with **training** and **development** data consisting of (*source*, *target*, *human post-edit*) triplets. This year, the two sets respectively comprise 18,000 and 1,000 instances, in which:

- The source (SRC) is an English sentence;
- The target (TGT) is a Marathi translation of the source produced by a generic, black-box NMT system unknown to participants. This multilingual NMT system (Ramesh et al., 2022) is based on the Transformer architecture (Vaswani et al., 2017) and is trained on a total of 49 million sentence pairs where the En-Mr parallel corpus is 4.5 million sentence pairs. This parallel data is generic and covers many domains, including the three domains covered by the evaluation setting of this year: healthcare, tourism/culture and general/news.
- The human post-edit (PE) is a manually-revised version of the target, which was produced by native Marathi speakers.

Also this year, a corpus of artificially-generated data has been released as additional training material. It consists of 2 million triplets derived from the *Anuvaad* en-mr parallel corpus[3]. The *Anuvaad* parallel corpus consists of data for 12 language pairs en-X, where X is 12 Indian languages, including Marathi. The English-Marathi data consists of 2.5 million parallel sentences. Specifically, the *source*, *target*, *post-edit* instances of this synthetic corpus are respectively obtained by combining: *i)* the original English source sentence from the *Anuvaad* corpus, *ii)* its automatic translation in Marathi[4], *iii)* the original Marathi target sentence from the *Anuvaad* corpus.

[2] Ethnologue-2022 - Ethnologue has been an active research project since 1951 which maintains online archives of recognized languages list, and their statistics.

[3] https://github.com/project-anuvaad/anuvaad-parallel-corpus

[4] from IndicTrans En-X Model (Ramesh et al., 2022)

Test data consisted of 1,000 (*source*, *target*) pairs, similar in nature to the corresponding elements in the train/dev sets (*i.e.,* same domains, same NMT system). The human post-edits of the target elements were left apart to measure APE systems' performance both with automatic metrics (TER, BLEU) and via manual assessments.

3.2 Metrics

In line with the previous rounds, also this year the plan was to evaluate the participating systems both by means of automatic metrics and, manually, via source-based direct human assessment (Graham et al., 2013). However, as discussed in Section 1, some issues in the manual evaluation procedure were later discovered. For this reason, the discussion of the evaluation results in Section 6 will only concentrate on the automatic metrics. Automatic evaluation was carried out after tokenizing the data using sacremoses[5] and then computing the distance between the automatic post-edits produced by each system for the target elements of the test set, and the human corrections of the same test items. Case-sensitive TER (Snover et al., 2006) and BLEU (Papineni et al., 2002) were respectively used as primary and secondary evaluation metrics. The official systems' ranking is hence 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. Automatic evaluation results are presented in Section 6.1.

3.3 Baseline

Also this year, the official baseline results were the TER and BLEU scores calculated by comparing the raw MT output with human post-edits. This corresponds to the score achieved by a "*do-nothing*" APE system that leaves all the test targets unmodified. For each submitted run, the statistical significance of performance differences with respect to the baseline was calculated with the bootstrap test (Koehn, 2004).

4 Complexity Indicators

To get an idea of the difficulty of the task, in previous rounds, we focused on three aspects of the released data, which provided us with information about the possibility of learning useful correction patterns during training and successfully applying them at test time. These are: *i)* repetition rate, *ii)* MT quality, and *iii)* TER distribution in the test set. For the sake of comparison across the eight rounds of the APE task (2015–2022), Table 1 reports, for each dataset, information about the first two aspects. The third one, instead, will be discussed by referring to Figure 1.

4.1 Repetition Rate

The repetition rate (RR), measures the repetitiveness inside a text by looking at the rate of non-singleton n-gram types (n=1...4) and combining them using the geometric mean. Larger values indicate a higher text repetitiveness that may suggest a higher chance of learning from the training set correction patterns that are also applicable to the test set. However, over the years, the influence of repetition rate in the data on system performance was found to be marginal.[8]

Looking at the data released this year, the very low RR values (*i.e.,* 1.46, 0.89, and 0.72 respectively for the SRC, TGT and PE elements) seem to confirm that repetition rate is a scarcely reliable complexity indicator. On one side, these values are close to those observed in rounds were the top-ranked submissions achieved both very large (2020) and very small (2021) gains over the baseline. On the other side, the best result for this year is close to the best results obtained, in previous rounds, on data featuring considerably higher repetition rates (2016, 2017). This suggests that other complexity factors may provide more reliable insights about the difficulty of the task, possibly with an additive effect, still to be fully understood, given by repetition rate.

4.2 MT Quality

Another possible complexity indicator is MT quality, that is the initial quality of the machine-translated (TGT) texts to be corrected. We measure it by computing, the TER (↓) and BLEU (↑) scores (Basel. TER/BLEU rows in Table 1) using the human post-edits as reference. In principle, higher quality of the original translations leaves the APE systems with smaller room for improvement since they have, at the same time, less to learn during

[5] https://pypi.org/project/sacremoses/

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

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

[8] The analyses carried out over the years produced mixed outcomes, with impressive final results obtained in spite of low repetition rates (Chatterjee et al., 2020) and vice-versa (Chatterjee et al., 2018, 2019; Akhbardeh et al., 2021).

	Lang.	Domain	MT type	RR_SRC	RR_TGT	RR_PE	Basel. BLEU	Basel. TER	δ TER
2015	en-es	News	PBSMT	2.9	3.31	3.08	n/a	23.84	+0.31
2016	en-de	IT	PBSMT	6.62	8.84	8.24	62.11	24.76	-3.24
2017	en-de	IT	PBSMT	7.22	9.53	8.95	62.49	24.48	-4.88
2017	de-en	Medical	PBSMT	5.22	6.84	6.29	79.54	15.55	-0.26
2018	en-de	IT	PBSMT	7.14	9.47	8.93	62.99	24.24	-6.24
2018	en-de	IT	NMT	7.11	9.44	8.94	74.73	16.84	-0.38
2019	en-de	IT	NMT	7.11	9.44	8.94	74.73	16.84	-0.78
2019	en-ru	IT	NMT	18.25	14.78	13.24	76.20	16.16	+0.43
2020	en-de	Wiki	NMT	0.65	0.82	0.66	50.21	31.56	-11.35
2020	en-zh	Wiki	NMT	0.81	1.27	1.2	23.12	59.49	-12.13
2021	en-de	Wiki	NMT	0.73	0.78	0.76	71.07	18.05	-0.77
2022	en-mr	healthcare/ tourism/news	NMT	1.46	0.89	0.72	67.55	20.28	-3.49

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 column (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.

training and less to correct at the test stage. On one side, 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 *i)* drastically reduce the number of corrections required and the applicability of the learned patterns, and *ii)* increase the chance of introducing errors, especially when post-editing near-perfect TGTs. The findings of all previous rounds of the task support this observation, which is corroborated by the high correlation (>0.83) between the initial MT quality ("Basel. TER" in Table 1) and the TER difference between the baseline and the top-ranked submission ("δ TER" in Table 1).

As discussed in Section 6, this year seems to confirm the trends observed in the past, albeit with a less evident match. The quality of the initial translations (20.28 TER / 67.55 BLEU) places this round among those of medium-high difficulty (20.0<TER<25.0) for which, except in one case (2015[9]), the performance gains obtained by the top-ranked submissions fall in the range -3.2<δ TER<-6.2. The δ TER of this year (-3.49) also falls in this range, confirming the correlation between the quality of the initial translations and the actual potential of APE.

4.3 TER Distribution

A third complexity indicator is the TER distribution (computed against human references) for the translations present in the test sets. Although TER distribution and MT quality can be seen as two sides of the same coin, it's worth remarking that, even at the same level of overall quality, more/less peaked distributions can result in very different testing conditions. Indeed, as shown by previous analyses, harder rounds of the task were typically characterized by TER distributions particularly skewed towards low values (*i.e.,* a larger percentage of test items having a TER between 0 and 10). On one side, the higher the proportion of (near-)perfect test instances requiring few edits or no corrections at all, the higher the probability that APE systems will perform unnecessary corrections penalized by automatic evaluation metrics. On the other side, less skewed distributions can be expected to be easier to handle as they give automatic systems larger room for improvement (*i.e.,* more test items requiring - at least minimal - revision). In the lack of more focused analyses on this aspect, we can hypothesize that in ideal conditions from the APE standpoint,

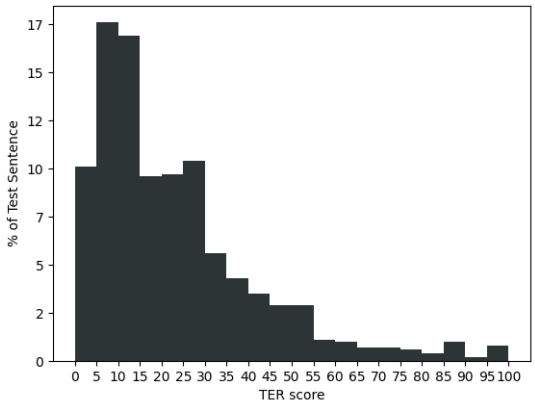

Figure 1: TER distribution in the APE 2022 English-Marathi test set.

[9]The 2015 round is the one in which the APE task was launched. It is somehow an exception being one of the two cases in which none of the participants managed to beat the *do-nothing* baseline (the other one was the 2019 sub-task on English-Russian, also exceptional in the choice of the target language).

ID	Participating team
IITB	Computation for Indian Language Technology - IIT Bombay, India (Deoghare and Bhattacharyya, 2022)
IIIT-Lucknow	IDIAP Research Institute, Switzerland
LUL	Samsung Research and Communication University of China, China (Xiaoying et al., 2022)

Table 2: Participants in the WMT22 Automatic Post-Editing task.

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.[10]

Also, with respect to this complexity indicator, the APE 2022 test set can be considered of medium-high difficulty compared to the past rounds. As shown in Figure 1, the TER distribution is quite skewed towards lower values (about 45% of the samples fall in the 15<TER<45 interval) but only 10% of the items can be considered as perfect or near-perfect translations (*i.e.,* 0<TER<5). These values are lower compared to those observed in the test data of harder rounds and higher compared to those observed in the test data of easier rounds.[11] All in all, the improvements over the baseline observed this year for two of the three participating systems (respectively -3.49 and -1.22 TER for the top-ranked and the second-best one) seem to confirm the correlation between TER distribution and task difficulty. However, weighing and understanding the actual contribution of TER distribution and MT quality, together with the possible additive effect of RR, remains a topic for more focused future research.

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

[11]Although the final results are not comparable due to the different evaluation settings (*i.e.,* different target languages and data domains), the findings from the last two rounds of the APE task provide good examples. In the 2021 round (English-German), where the top submission achieved a small TER reduction compared to the baseline (-0.77), more than 35% of the test instances featured a TER between 0 and 5 and almost 50% of them had 0<TER<10. In contrast, in the 2020 round (English-Chinese) where the top submission achieved the largest baseline improvement ever observed (-12.13), less than 1% of the test samples had 0<TER<5 and ∼89% of them had 40<TER<85.

5 Submissions

As shown in Table 2, this year we received submissions from three teams. Two of them (IIIT-Lucknow and LUL) submitted two runs, while the third one (IITB) participated with only one submission. The main characteristics of two of the three participating systems are summarized below.[12]

Samsung Research and Communication University of China (LUL). This team participated with a Transformer-based system built using fairseq (Ott et al., 2019). Their submissions are characterized by two main aspects: data augmentation and the use of a mixture of experts' approach (Jacobs et al., 1991). Data augmentation is pursued by generating synthetic triplets by means of both an in-house MT system and an external system (Google Translate). The former is used to translate text drawn from several resources, while the latter is used to back-translate the post-edits in the APE training set. The resulting material is combined in different ways so as to obtain different data sets for model fine-tuning. The mixture of experts' approach exploits three domain-specific adapters (Bapna and Firat, 2019; Pham et al., 2020), which are added to the decoder of the base APE model. At inference time, a classifier (added after the encoder) is used to decide which adapter has to be activated.

Computation for Indian Language Technology - IIT Bombay (IITB). This team participated with a Transformer-based system. It exploits a multi-source approach similar to the one in (Chatterjee et al., 2017), with two separate encoders to generate representations for SRC, MT and one decoder. The model is trained with a curriculum learning strategy similar to the one applied by the 2021 winning system (Oh et al., 2021). This is done by first incrementally using out-/in-domain synthetic data (*i.e.,* those released to participants and additional

[12]The IIIT-Lucknow did not produce a system description paper and is left out of our analysis.

ones generated via MT) and then by fine-tuning the model on the real APE data. To ensure the quality of the training material, the LaBSE technique (Language-agnostic BERT sentence embedding) by Feng et al. (2022) is used to filter out low-quality synthetic triplets. To reduce over-correction, a sentence-level quality estimation system trained on the WMT-22 QE English-Marathi sub-task is used to select the final output between an original translation and the corresponding (corrected) version generated by the APE model.

6 Results

6.1 Automatic Evaluation

Participants' results are shown in Table 3. The submitted runs are ranked based on the average TER (case-sensitive) computed using human post-edits of the MT segments as a reference, which is the APE task's primary evaluation metric. We also report the BLEU score, computed using the same references, which represents our secondary evaluation metric.

As it can be seen from the table, the two rankings are coherent: the top submission (16.79 TER, 72.92 BLEU) is the same, and the top three systems outperform by a large margin ($\sim$1 TER and $\sim$2 BLEU scores) the *do nothing* baseline, both in term of BLEU and TER score. These systems are statistically better than the baseline. This is indeed an interesting result showing the effectiveness of the APE systems and confirming their capability of profitably leveraging additional and external resources compared to the MT system.

Looking at relationships between the primary and contrastive submissions (IIT and LUL), the contrastive system shows slightly better performance of the primary submission in one case (LUL). This highlights the difficulty to select the best configuration during system development and indirectly confirms the difficulty to handle APE data characterized by high MT quality, and TER distribution skewed towards perfect/near-perfect translations.

6.2 Systems' Behaviour

Modified, improved and deteriorated sentences. To better understand the behaviour of each APE system, we now turn an eye toward the changes made by each system to the test instances. To this aim, Table 4 shows, for each submitted run, 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 for which improvement/deterioration is observed). It's worth noting that, as in the previous rounds, 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 any TER variations. This grey area, for which quality improvement/degradation can not be automatically assessed, would contribute to motivating the integration of human assessments, as done previously.

As it can be seen from the table and similarly to last year's edition, the top systems have been quite conservative in applying their edits by modifying a limited percentage of sentences ($\sim$50% on average, 45.2 for the top submission). Considering the TER distribution where a large number of samples lay in the 15<TER<45 interval, there is the possibility of substantially changing the MT outputs to achieve better performance. This limited number of edits is unexpected and similar to more difficult test sets with more skewed TER distributions toward near-perfect translations. However, systems' final scores are inversely proportional to their aggressiveness showing that limiting the APE edits and carefully selecting them is the right strategy toward significant improvements in quality.

Precision-wise, this year's systems reached 63.9 (in 2021 it was 51.12 and 58.0 in 2020) on average with the best run peaking at 69.49 (vs 53.96 in 2021 and 69.0 in 2020). It is important to note that the average value is significantly affected by the low-performing systems having a precision close to 0. Looking at the percentage of improved (55.6 on average, 63.49 for the top submission) and deteriorated (31.2 on average, 27.87 for the winning system) sentences, the results confirm the capability of the top systems to minimize the wrong changes. Compared to the last editions, the percentage of the improved sentences is among the largest ones achieved by the all-time submitted APE systems.

Edit operations. Similar to previous rounds, we analysed systems' behaviour also in terms of the distribution of edit operations (insertions, deletions, substitutions and shifts) done by each system. This fine-grained analysis of how systems corrected the test set instances is obtained by computing the TER between the original MT output and the output of each primary submission taken as a reference. Sim-

		TER	BLEU
en-mr	IITB_APE_QE_combined_PRIMARY.tsv	**16.79**	**72.92**
	LUL_HyperAug_Adaptor_CONTRASTIVE	**19.06**	**69.96**
	LUL_HyperAug_Finetune_PRIMARY	**19.36**	**69.66**
	baseline (MT)	20.28	67.55
	IIIT-Lucknow_adversia-machine-translation_PRIMARY.txt	57.14	23.43
	IIIT-Lucknow_adversia-machine-translation_CONTRASTIVE.txt	99.81	3.16

Table 3: Results for the WMT22 APE English-Marathi shared task – average TER (↓), BLEU score (↑) Statistically significant improvements over the baseline are marked in **bold**.

Systems	Modified	Improved	Deteriorated	Prec.
IITB_APE_QE_combined_PRIMARY	452 (45.2%)	287 (63.49%)	126 (27.87%)	69.49
LUL_HyperAug_Adaptor_CONTRASTIVE	491 (49.1%)	261 (53.15%)	150 (30.54%)	63.5
LUL_HyperAug_Finetune_PRIMARY	537 (53.7%)	269 (50.09%)	189 (35.19%)	58.73
IIIT-Lucknow_adversia-machine-translation_PRIMARY	999 (99.9%)	46 (0.46%)	929 (92.99%)	0.47
IIIT-Lucknow_adversia-machine-translation_CONTRAS.	1000 (100%)	9 (0.09%)	987 (98.7%)	0.09
Average	69.6 (49.3)	31.4 (55.6)	57.0 (31.2)	38.4 (63.9)

Table 4: Number (raw and proportion) of test sentences modified, improved and deteriorated by each run submitted to the APE 2022 English-Marathi sub-task. The "Prec." column shows systems' precision as the ratio between the number of improved sentences and the number of modified instances for which improvement/deterioration is observed (*i.e.,* Improved + Deteriorated).

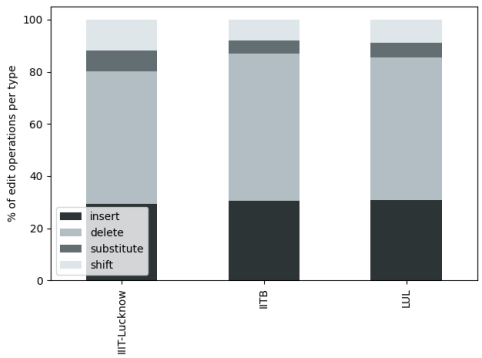

Figure 2: Distribution of edit operations (insertions, deletions, substitutions and shifts) performed by the three primary submissions to the WMT22 APE English-Marathi shared task.

ilar to last year, differences in systems' behaviour are minimal. All of them are characterised by a large number of deletions (∼55.0% on average), followed by insertions (∼30%), shifts (∼10%) and substitutions (∼6%). The system that seems to have a slightly different distribution is IIT-Lucknow resulting in more shifts and substitutions, but these differences are barely visible. Although this year's test set turned out to be simpler than last year (less shewed TER distribution and higher TER), the edit operations are very similar to last year's with a small difference in the number of deletions (65% last year, 55% this year) and insertions (19.2% vs 30%). These variations may depend on the new data, target language and MT system. More thorough future investigations would be needed to find clear explanations for these observations.

7 Conclusion

The 8^{th} round of the shared task on Automatic Post-Editing at WMT was characterized by two main factors of novelty: the language pair (English-Marathi) and the domain of the released data (a mix covering healthcare, tourism, and general/news). Apart from this, the overall setting was the same as in previous recent rounds, in which participating systems had to automatically correct the output of a generic neural MT system, being evaluated with the TER (primary) and BLEU (secondary) automatic metrics. In continuity with the past, also human evaluation via source-based direct assessment was carried out, but it is not discussed in this report due to its unreliable outcomes. In terms of the three complexity indicators discussed in Section 4 (repetition rate, original MT quality and TER distribution), the difficulty of this round falls in a medium-high range. This is reflected by the performance of the systems submitted by the three participating teams: two of them were indeed able to improve over the *do-nothing* baseline with (statistically significant) error reductions up to -3.49 TER points (+5.37 BLEU). Although these results are not comparable with those from previous years due to the different language/domain testing conditions, the observed improvements in the new language direction confirm the viability of APE for downstream improvements of "black-box" MT systems whose inner workings are not accessible.

Acknowledgments

We would like to thank the translation agencies Techliebe, Shri Samarth Krupa Language Solutions, Zibanka, and Desicrew, who helped post-edit the dataset for the English-Marathi language pair.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Aakash Banerjee, Aditya Jain, Shivam Mhaskar, Sourabh Dattatray Deoghare, Aman Sehgal, and Pushpak Bhattacharyya. 2021. Neural machine translation in low-resource setting: a case study in english-marathi pair. In *Proceedings of the 18th Biennial Machine Translation Summit (Volume 1: Research Track)*, pages 35–47.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1538–1548, Hong Kong, China. Association for Computational Linguistics.

Rajen Chatterjee, M. Amin Farajian, Matteo Negri, Marco Turchi, Ankit Srivastava, and Santanu Pal. 2017. Multi-source neural automatic post-editing: Fbk's participation in the wmt 2017 ape shared task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 630–638, Copenhagen, Denmark. 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, Markus Freitag, Matteo Negri, and Marco Turchi. 2020. Findings of the WMT 2020 shared task on automatic post-editing. In *Proceedings of the Fifth Conference on Machine Translation*, pages 646–659, Online. Association for Computational Linguistics.

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

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.

Sourabh Deoghare and Pushpak Bhattacharyya. 2022. Iit bombay's wmt22 automatic post-editing shared task submission. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. 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.

R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. 1991. Adaptive mixtures of local experts. *Neural Computation*, 3:79–87.

Jyotsana Khatri, Rudra Murthy, Tamali Banerjee, and Pushpak Bhattacharyya. 2021. Simple measures of bridging lexical divergence help unsupervised neural machine translation for low-resource languages. *Machine Translation*, 35(4):711–744.

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

Shinhyeok Oh, Sion Jang, Hu Xu, Shounan An, and Insoo Oh. 2021. Netmarble AI center's WMT21 automatic post-editing shared task submission. In *Proceedings of the Sixth Conference on Machine Translation*, pages 307–314, Online. Association for Computational Linguistics.

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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Stroudsburg, PA, USA. Association for Computational Linguistics.

Minh Quang Pham, Josep Maria Crego, François Yvon, and Jean Senellart. 2020. A study of residual adapters for multi-domain neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 617–628, Online. Association for Computational Linguistics.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

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.

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.

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

Huang Xiaoying, Lou Xingrui, Zhang Fan, and Tu Mei. 2022. Lul's wmt22 automatic post-editing shared task submission. In *Proceedings of the Seventh Conference on Machine Translation*, Abu Dhabi. Association for Computational Linguistics.