Findings of the WMT 2023 Shared Task on Automatic Post-Editing

Pushpak Bhattacharyya
IIT Bombay
pb@cse.iitb.ac.in

Rajen Chatterjee
Apple Inc.
rajen_c@apple.com

Markus Freitag
Google
freitag@google.com

Diptesh Kanojia
University of Surrey
d.kanojia@surrey.ac.uk

Matteo Negri
Fondazione Bruno Kessler
negri@fbk.eu

Marco Turchi
Zoom Video Communications
marco.turchi@zoom.us

Abstract

We present the results from the 9^{th} round of the WMT shared task on MT Automatic Post-Editing, which consists of automatically correcting the output of a "black-box" machine translation system by learning from human corrections. Like last year, the task focused on English→Marathi, with data coming from multiple domains (healthcare, tourism, and general/news). Despite the consistent task framework, this year's data proved to be extremely challenging. As a matter of fact, none of the official submissions from the participating teams succeeded in improving the quality of the already high-level initial translations (with baseline TER and BLEU scores of 26.6 and 70.66, respectively). Only one run, accepted as a "late" submission, achieved automatic evaluation scores that exceeded the baseline.

1 Introduction

This paper presents the results of the 9^{th} round of the WMT task on MT Automatic Post-Editing (APE). The task involves the automatic correction of the output generated by 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 consistent with that of all previous rounds. In this formulation, the challenge revolves around fixing errors in English documents that have been automatically translated by a state-of-the-art, non-domain-adapted neural MT (NMT) system unknown to the participants. In continuity with last year's round, the evaluation focused on English→Marathi,[1] with training/dev/test data selected from a mix of domains, namely- healthcare, tourism, and general/news (see Section 3).

Three teams participated in the task by submitting a total of four runs for the final evaluation (see Section 4).[2] However, while only two out of the three participants were able to submit their runs on time, the one remaining submission arrived with a two-month delay. This led us to categorize it as a late (therefore, unofficial) submission for the sake of fairness to the other participants.

For all the teams, the task posed significant challenges primarily due to the high average quality of the initial translations slated for post-editing (26.6 TER / 70.66 BLEU / 79.78 chrF). This challenge was compounded by the substantial imbalance in distribution between near-perfect translations (approximately 40% of the total) and those necessitating extensive revisions (approximately 20%). As a consequence, none of the official runs was able to improve over the baseline in terms of the task's automatic evaluation metrics (Section 5.1), with the best run achieving results (27.73 TER / 69.03 BLEU / 78.64 chrF) that highlight a slight quality degradation compared to the original, untouched NMT outputs that represent our baseline. For the sake of completeness, we report that the late submission achieved a slight improvement over the baseline, attested by TER, BLEU, and chrF scores of 25.74, 71.27, and 80.41, respectively. The results computed by means of automatic evaluation metrics were also confirmed by our human evaluation based on direct assessment (Section 5.2).

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:

- Enhance MT output by harnessing information that is not available to the decoder or by conducting deeper text analysis, which may

[1] Marathi is an Indo-Aryan language predominantly spoken by Marathi people in the Indian state of Maharashtra.

[2] A fourth participant withdrew the submitted run, which was affected by major errors in the generated outputs.

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 672–681
December 6–7, 2023.

be prohibitively expensive during the decoding phase.

- Address systematic errors stemming from an MT system whose decoding process is inaccessible for focused modifications.
- Provide professional translators with improved MT output quality, thereby reducing the need for subsequent human post-editing.
- Tailor the output of a general-purpose MT system to align with the lexicon and style requirements of a specific application domain.

This 9^{th} round of the WMT APE shared task kept the same overall evaluation setting of the previous eight 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. For the second year in a row, the selected language pair was English-Marathi (with Marathi as the target language for post-editing). Training, development and test data were drawn from the following three domains: healthcare, tourism, and general/news.

3 Data, Metrics, Baseline

3.1 Data

In continuity with last year, the selected language pair is 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.[3] Marathi is a known agglutinative language and presents various challenges to machine translation compared to its other Indian counterparts (Khatri et al., 2021; Banerjee et al., 2021). Moreover, the English-Marathi language pair is considered low-resource compared to English-Hindi/Bengali/Malayalam (Ramesh et al., 2022), despite having more native speakers worldwide.

The **training** and **development** datasets supplied to the participants remain consistent with those employed in the 2022 iteration of the task. These datasets consist of 18,000 and 1,000 (*source*, *target*, *human post-edit*) triplets, wherein:

- 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 APE 2023 test set, namely-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.

We provide the same corpus of artificially generated data as additional training material from the last round. It consists of 2 million triplets derived from the *Anuvaad* en-mr parallel corpus.[4] The *Anuvaad* parallel corpus consists of data for 12 en-X language pairs, where X comprises 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 into Marathi,[5] *iii)* the original Marathi target sentence from the *Anuvaad* corpus.

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, chrF) and via human evaluation.

3.2 Metrics

The participating systems were evaluated both by means of automatic metrics and manually. Automatic evaluation (Section 5.1) was carried out after tokenizing the data with *sacremoses*[6], by computing the distance between the automatic post-edits produced by each system for the target elements of

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

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

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

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

the test set, and the human corrections of the same test items.

The official systems' ranking is based on the average (case-sensitive) TER (Snover et al., 2006) calculated on the test set by using the TERcom[7] software: lower average TER scores correspond to higher ranks. As additional performance indicators, BLEU (Papineni et al., 2002) and chrF (Popović, 2015) were computed[8]. The human evaluation (Section 5.2) was conducted via source-based direct human assessment (Graham et al., 2013a).

3.3 Baseline

The official baseline results were the TER/BLEU/chrF 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.

4 Submissions

As shown in Table 1, this year, we received submissions from three teams, one of which submitted their run with a two-month delay that motivates its categorization as a late submission[9]. The main characteristics of the participating systems are summarized below.

Korea Advanced Institute of Science and Technology (kaistai). This team participated with a system inspired by the recent surge of large language models (LLMs) that have been successfully applied to a variety of language generation tasks. Their goal was to verify whether LLMs could perform the APE task through prompting. To this aim, they used gpt-3.5-turbo with specific prompts designed to generate either (a) post-edits or (b) post-edits along with the rationales behind them. While the results of preliminary evaluations based on COMET suggested the viability of the approach for medium-/high-resource language pairs, they also highlighted that the often radical changes produced by LLMs can potentially be penalized by more strict reference-based evaluations based on BLEU, TER, or chrF.

Korea University (KU_UPs). The participation of this team was centred on data filtering techniques. With a focus on removing potentially harmful material from a model training perspective, the proposed method concentrates on eliminating the two extremes of the training data distribution: the (near-)perfect MT outputs on one side, and those that require complete rewriting on the other. According to preliminary experiments carried out on previous APE datasets (WMT2020/2021/2022), data selection driven by TER and COMET yields better performance when the outlier instances requiring excessive post-editing are removed from the training. On this basis, the submitted APE system was built by training the multilingual M2M100-418M model (Fan et al., 2021).

Huawei Translation Service Center and Xiamen University School of Informatics (HW_TSC). – late submission – This team participated with a Transformer-based system pre-trained on the provided synthetic APE data and then fine-tuned on the real APE data augmented via automatic translation (with Google Translate run on the post-edits in the training set) and by integrating En-Mr parallel sentences from FLORES-200 (NLLB Team et al., 2022). R-Drop (Liang et al., 2021), which regularizes the training inconsistency induced by dropout, is used to mitigate overfitting during the training phase. A sentence-level Quality Estimation system is also used to select the most appropriate output, choosing between the original translation and the corresponding APE-generated output.

5 Results

5.1 Automatic Evaluation

Automatic evaluation results are shown in Table 2. 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. To provide a broader view of systems' performance, BLEU and chrF results computed using the same references are also reported.

As can be seen from the table, the three rankings coherently show that the best official submission (by the KU_UPS team, which achieved scores of 27.73 TER, 69.03 BLEU, and 78.64 chrF) outperforms the others. None of them, however, was able to improve the quality of the original translations (*i.e.* the *do nothing* baseline), differently from

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

[8] chrF was computed using SacreBLEU https://pypi.org/project/sacrebleu/(version 2.3.0)

[9] A fourth participating team retracted their submitted run due to errors in the generated outputs that significantly affected their final results.

ID	Participating team
kaistai	Korea Advanced Institute of Science and Technology, South Korea
KU_UPs	Korea University, South Korea (Moon et al., 2023)
HW_TSC	Huawei Translation Service Center & Xiamen University School of Informatics, China (Yu et al., 2023)

Table 1: Participants in the WMT23 Automatic Post-Editing task.

		TER	BLEU	CHRF
en-mr	HW-TSC_HW_1_PRIMARY.txt†	25.74	71.27	80.41
	baseline (MT)	26.60	70.66	79.78
	KU_UPs-filtered4-PRIMARY.tsv	27.73	69.03	78.64
	kaistai_prompt-wo-cot_contrastive	54.59	40.97	67.24
	kaistai_prompt-w-cot_primary	58.55	31.63	61.61

Table 2: NEWResults for the WMT23 APE English-Marathi shared task – average TER (↓), BLEU (↑), chrF (↑). Gray† indicates a late submission, which was received after the conclusion of this year's human evaluation and, consequently, is not discussed in Section 5.2.

the slightly better outputs of the late submission by HW_TSC. This prompts further analyses to explore the underlying reasons for this unexpected outcome. We do this in two ways: 1) by giving a closer look at systems' behaviour (Section 5.1.1), in order to spot trends in their post-editing strategies; 2) by analysing the task's inherent level of difficulty (Section 5.1.2) in terms of the possibility to learn valuable correction patterns from the training data and effectively apply them to the supplied test set.

5.1.1 Analysis: Systems' Behaviour

Modified, improved and deteriorated sentences To gain a first insight into the behaviour of participating systems, Table 3 provides an overview of each submitted run, detailing the number of modified, improved, and deteriorated sentences, along with the systems' overall precision (*i.e.,* the ratio of improved sentences to the total count of modified instances where improvement or deterioration is observed). It's worth noting that each system has modified a much higher number of sentences than the combined total of improved and deteriorated ones. This discrepancy accounts for modified sentences in which the corrections do not result in any variations in TER. This "grey area", where the automatic assessment of quality improvement or degradation is not feasible, underscores the importance of including human evaluation for a comprehensive assessment of systems' performance (see Section 5.2). As can be seen from the table, and in line with the findings from previous rounds, conservative post-editing seems to yield better results compared to the adoption of aggressive strategies. The difference between the top-ranked system and the other submitted runs is indeed evident when we look at the proportion of modified test sentences (37.4%[10] vs $\geq 93.1\%$), indicating that limiting the applied edits to the strictly necessary ones remains the main challenge to achieve significant quality improvements. While this outcome may be influenced by the reference-based automatic evaluation framework employed (as it penalizes correct edit operations that deviate from those presented in the reference), it is noteworthy that the results of the manual evaluation, as detailed in Section 5.2, align with this observation.

Another observation is that precision is certainly the other key factor in achieving good APE results. Besides being much more conservative, the best submission stems, in fact, for a higher precision in selecting the edit operations to be applied (48.11[11] vs ≤ 21.00). Also, this finding aligns with the outcomes of previous rounds, in which the winning system consistently exhibited the highest precision. Notably, the precision of this year's official submissions (averaging 29.62) is significantly lower than the values observed in previous rounds (*e.g.,* 69.0, 53.96, 69.49 for the top-ranked system in 2020, 2021, and 2022, respectively).[12] This difference in precision may well explain why none of them were able to improve upon the baseline results.

Edit operations Further indications about the system's behaviour can be drawn from a more fine-

[10] Which drops to 24.4% for the late submission.

[11] Further increased to 51.89 for the late submission.

[12] This holds even if we include the late submission in the computation, with an average precision that slightly grows to 35.19.

Systems	Modified	Improved	Deteriorated	Prec.
HW-TSC_HW_1_PRIMARY.txt†	244 (24.4%)	110 (45.08%)	102 (41.8%)	51.89
KU_UPs-filtered4-PRIMARY.tsv	374 (37.4%)	153 (40.9%)	165 (44.11%)	48.11
kaistai_prompt-wo-cot_contrastive	931 (93.1%)	187 (20.08%)	709 (76.15%)	20.87
kaistai_prompt-w-cot_primary	989 (98.9%)	193 (19.51%)	777 (78.56%)	19.89
Average	76.46%	26.83%	66.27%	29.62

Table 3: Number (raw and proportion) of test sentences modified, improved and deteriorated by each run submitted to the APE 2023 English-Marathi subtask. The "Prec." column shows systems' precision as the ratio between the number of improved sentences and the total number of modified instances for which improvement or deterioration can be assessed in terms of TER variations. Average values considering only the three official submissions.

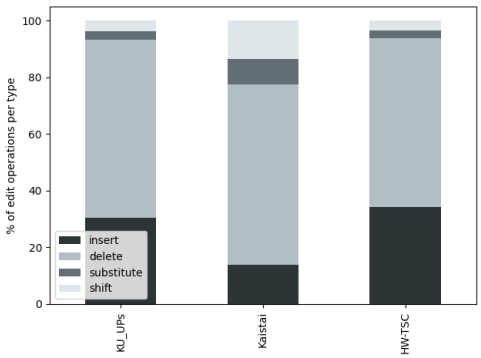

Figure 1: Distribution of edit operations (insertions, deletions, substitutions and shifts) performed by the three primary submissions to the WMT23 APE English-Marathi shared task.

grained analysis of the distribution of their edit operations (insertions, deletions, substitutions, and shifts). Such distribution is obtained by computing the TER between the original MT output and the output of each primary submission, taken as a reference. As shown in Figure 1, although the overall behaviour of the systems is similar, some differences are noticeable. Indeed, in line with previous rounds, they all exhibit a high percentage of deletions, followed by insertions, substitutions and shifts. However, for the best official submission,[13] the percentage of the latter two types of operations is minimal (2.9% substitutions and 3.67% shifts) and balanced by a less skewed distribution of insertions (30.52%) and deletions (62.91%). Especially the comparatively higher proportion of more "radical" (*i.e.,* structural) modifications applied by the worse system (13.43% shifts), which again suggests its lower conservativeness, can account for its lower automatic evaluation scores.

[13]Note, however, that the same consideration also applies for the late submission.

5.1.2 Analysis: Complexity Indicators

While systems' behaviour is influenced by implementation and architectural choices on the one hand, it also depends on the data used for training, development, and evaluation on the other. Therefore, looking at the intrinsic difficulty of the task from a data perspective is also crucial for interpreting the observed performance of the systems. To delve into this aspect, we concentrate on the possibility of learning useful correction patterns during training and successfully applying them at test time. We analyse such a possibility in terms of three indicators, namely: *i)* repetition rate, *ii)* MT quality, and *iii)* TER distribution in the test set. For the sake of comparison across the nine rounds of the APE task (2015–2023), Table 4 reports, for each dataset, information about the first two aspects. The third one, instead, will be discussed by referring to Figure 2.

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. Higher RR values indicate greater text repetitiveness, which may imply an increased likelihood of learning correction patterns from the training set that are also applicable to the test set. As shown in the last row of Table 4, the RR values for the SRC, TGT, and PE elements (averaged across the training, development, and test sets) are relatively low. Furthermore, upon closer examination, Table 5 reveals a non-negligible difference between the RR values of the SRC, TGT, and PE elements in the training set compared to the corresponding values calculated on the test set. This difference is particularly pronounced for the PE sentences, where the RR is more than two times higher. Although the reported RR values can be considered indicative of a challenging task, it is important to

	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	health/tourism/news	NMT	1.46	0.89	0.72	67.55	20.28	-3.49
2023	en-mr	health/tourism/news	NMT	1.85	1.24	1.12	70.66	26.60	+1.13

Table 4: 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 official submission.

Data	RR
Train_src	1.55
Train_mt	1.03
Train_pe	0.81
Dev_src	1.4
Dev_mt	0.8
Dev_pe	0.64
Test_src	2.6
Test_mt	1.9
Test_pe	1.91

Table 5: Repetition Rate (RR) values of source (src), target translation (mt) and post-edited translation (pe) elements in the APE 2023 training, development and test sets.

note that the top-ranked submissions in previous rounds (e.g. in 2022 and 2020) were able to achieve significant improvements over the baseline despite similar RR values (with δ TER values of -3.49 and -12.13, respectively). This variability reinforces the findings from previous rounds, emphasizing that RR alone is insufficient as a complexity indicator. Rather, it underscores the significance of examining its interaction with other indicators and its potential cumulative impact on them.

MT Quality As emphasized by the findings from all previous rounds of the task, a more reliable indicator of complexity is the quality of the machine-translated (TGT) texts that require correction. We assess this quality by computing TER (↓) and BLEU (↑) scores (shown in the Basel. TER/BLEU columns in Table 4), using the human post-edits as references.[14] In principle, higher-quality original translations leave less room for improvement to APE systems, which have at the same time fewer errors to learn from during training and fewer corrections to make at test time. On one side, indeed, training on good (or near-perfect) automatic translations can significantly reduce the number of learned correction patterns. On the other side, testing on similarly high-quality translations can have two effects: *i)* it reduces the number of corrections required and the applicability of learned patterns, and *ii)* it increases the risk of introducing errors, especially when post-editing near-perfect TGTs. This observation is supported by the strong correlation (>0.83) between the initial MT quality ("Basel. TER" in Table 4) and the TER difference between the baseline and the top-ranked submission ("δ TER" in Table 4) previously reported in the analysis of the 2015-2022 rounds by Bhattacharyya et al. (2022).

Looking at the baseline TER score, this year's test data look for a comparatively lower difficulty for APE systems compared to most of the previous rounds, which in only 2 cases (*i.e.,* for the two languages covered in 2020) appear to be less challenging. Interestingly, however, when looking at the baseline BLEU score, the difficulty appears to be higher, with up to 6 previous test sets featuring translations of lower quality (hence easier to handle) compared to this year. The reasons for such differences deserve further investigation, which might shed light on the fact that, contrary to expectations, MT quality is less indicative of this year's task difficulty compared to previous rounds[15].

[14] Scores for the newly introduced chrF metric are not included in the table, as they would not be comparable with values from previous rounds where chrF was not considered.

[15] Considering this year's data, in fact, the correlation between "Basel. TER" and "δ TER" in Table 4 drops from >0.83

TER distribution in the test set Complementary to repetition rate and MT quality, the TER distribution (computed against human references) for the translations present in the test provides valuable insights for interpreting the results of this year's round of the task. While TER distribution and MT quality may appear to be closely related, it's important to note that, even at similar overall quality levels, more or less skewed distributions can create distinct testing conditions. Indeed, as shown by previous analyses (Bojar et al., 2017; Chatterjee et al., 2018, 2019, 2020; Akhbardeh et al., 2021; Bhattacharyya et al., 2022), more challenging rounds of the task were typically characterized by TER distributions heavily skewed toward lower values (*i.e.,* a larger percentage of test items having a TER between 0 and 10).

On one side, a higher proportion of (near-)perfect test instances, requiring minimal or no corrections, increases the likelihood that APE systems will make unnecessary edits, which will be penalized by automatic evaluation metrics. Conversely, less skewed distributions may be easier to handle, as they provide automatic systems with more opportunities for improvement, with a larger number of test instances necessitating revision. In the lack of more focused analyses on this aspect, we can hypothesize that under ideal conditions from the APE standpoint, the peak of the distribution would correspond to "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.[16] In light of the above observations, the APE 2023 test set can be considered as particularly challenging. As illustrated in Figure 2, the TER distribution exhibits a U-shaped (bimodal) pattern, characterized by two prominent peaks corresponding to the two most critical regions within the $0-100$ TER range. At one extreme, the first peak corresponds to the vast majority of test instances (about 45% of the total) that can be considered as perfect or near-perfect translations (*i.e.,* 0<TER<5), which implies a high chance of applying unnecessary corrections. At the other extreme, the second peak corresponds to a significant portion of test items (about 20%) that can be considered as too poor and unintelligible (*i.e.,* $95<TER<100$) to grant the safe application of any post-editing strategies. Although the remaining portion of the test set falls almost entirely in the range of "post-editable" outputs (*i.e.,* $10<TER<40$), its small size significantly reduces the potential for improvement through the APE process. Overall, this year's test set deviates significantly from all previous ones, where the TER distributions have never been characterized by such a pronounced bimodal pattern. In light of this, we can conclude that while, on the one hand, the repetition rate and machine translation quality do not provide sufficiently convincing insights to justify performance below the baseline for the official submissions, on the other hand, the TER distribution has posed a significant challenge for this year's participants.

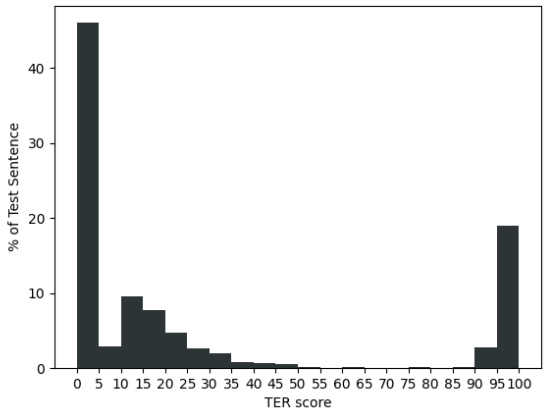

Figure 2: TER distribution in the APE 2023 English-Marathi test set.

5.2 Human Evaluation

We conducted a human evaluation of the primary system submissions to complement the automatic evaluations. However, this could be performed only for the official system submissions, as the late submission was received after the conclusion of the human assessments. This section discusses our evaluation procedure and the results obtained from it.

5.2.1 Evaluation Procedure

We provided annotation guidelines to professional translators who are native speakers of the target language. The same guidelines were also used to collect Indic language quality estimation shared task dataset (Zerva et al., 2022). The annotators provided a source-based direct assessment (DA) (Gra-

to 0.78.

[16]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.

	Avg DA	Avg z
test.pe	83.76	0.426
KU_UPs-filtered4-PRIMARY	66.56	−0.171
test.mt	65.86	−0.138
kaistai_prompt-w-cot_primary	64.79	−0.116

Table 6: Results for the human evaluation campaign for the En-Mr language pair. Systems ordered by DA score; systems within a cluster are considered tied; lines indicate clusters according to Wilcoxon rank-sum test p <0.05.

ham et al., 2013b; Cettolo et al., 2017; Bojar et al., 2018) score to each segment containing the source and the APE system output. We hired 4 translators to evaluate the two primary system submissions (KU_UP & KAISTAI), manually post-edited segments (*test.pe*), and the MT Output (*test.mt*). We chose to allocate an equal number of instances to each translator after shuffling, and only a single DA annotation was collected for each instance (Toral, 2020). Shuffling the instances before allocation helps prevent annotator bias towards a single system in the direct assessments.

The annotation guidelines provide a detailed description of potential adequacy and fluency-based errors based on which the translator could estimate the direct assessment score range. However, the translators were additionally instructed to prioritize adequacy errors and focus on assessing the semantic similarity between the source and the system output. The annotators manually entered the DA score between 0-100. The collected DA annotations were unshuffled based on the segment IDs, which were unknown to the translators. We expected the human post-editing to be of higher quality compared to APE system submissions and, consequently, better than the MT baseline.

5.2.2 Evaluation Results

We present the results obtained from the human evaluation campaign in Table 6. As expected, the human post-edited segments were rated the highest at 83.76 mean DA score. However, contrary to automatic evaluation, the submission by KU_UP was rated slightly better than the MT baseline (test.mt). But, the score difference in both cases-human and automatic evaluation, seems insignificant. Additionally, as per the Wilcoxon Rank-sum test, KU_UP and MT baseline score distributions seem to be in a cluster. In line with the automatic evaluation, the mean DA obtained by the submission from *kaistai* was rated the lowest at 64.79, lower than the MT baseline at 65.86. This submission utilizes LLMs to perform the APE task and raises a question on the viability of LLMs for APE when a low-resource language is concerned. LLMs are mostly fine-tuned and/or evaluated on task datasets in English (Hendrycks et al., 2020; Longpre et al., 2023), and there remain unanswered questions on their viability for complex and challenging multilingual tasks like APE. Owing to a challenging test set this year, our analysis highlights the difficulty posed by the task and implores us to consider a different setting in which the APE task can perhaps gain assistance through a translation quality signal. QE systems have been explored for assisting the APE task in a supervised multi-task scenario, which intuitively helps the model perform better at both tasks.

6 Conclusion

We presented the results from the 9^{th} shared task on Automatic Post-Editing at WMT. In continuity with the 2022 round, the task focused on the automatic correction of NMT outputs generated by a black-box English-Marathi system. The three participating systems were evaluated both automatically (with TER as the primary metric, BLEU, and ChrF) and manually. According to automatic evaluation results, only one (late) submission succeeded in outperforming the *do-nothing* baseline. The analysis of this year's data suggests that one of the main causes of difficulty might be the bimodal, U-shaped TER distribution of the test instances, which substantially differs from the test set distributions observed in all previous rounds (skewed but a pattern closer to normal). Our manual evaluation confirms the automatic evaluation outcomes and affirms the challenge posed by APE for the current approaches. We observe that one of the systems performs quite close to the MT baseline while the other performs well below the same. Additionally, the lack of multilingual datasets in LLM training/benchmarking raises a question on the viability of performing challenging multilingual tasks like APE. All in all, these findings advocate for further research on this challenging problem, which, far from being solved, this year revealed new nuances in terms of difficulty. Next year, we plan to introduce two new low-resource language pair datasets for the APE task. Future developments will also

likely include a re-definition of some aspects of the evaluation settings, which have remained relatively stable over the years. For instance, the set of automatic evaluation metrics will likely be reconsidered and expanded so as to include more semantics-oriented measures, with an eye on the advent of large language models increasingly adopted also for APE.

Acknowledgements

We would like to thank Zibanka Media Services Pvt. Ltd., and Techliebe, who worked with us to create the APE datasets for English-Marathi.

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.

Pushpak Bhattacharyya, Rajen Chatterjee, Markus Freitag, Diptesh Kanojia, Matteo Negri, and Marco Turchi. 2022. Findings of the WMT 2022 shared task on automatic post-editing. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 109–117, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. 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: Shared Task Papers*, pages 272–303, 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 *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 2–14, Tokyo, Japan. International Workshop on Spoken Language Translation.

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.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1).

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

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013b. 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.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

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.

Xiaobo* Liang, Lijun* Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. 2021. R-drop: Regularized dropout for neural networks. In *Proceedings of NeurIPS*.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*.

Hyeonseok Moon, Seugnjun Lee, Chanjun Park, Jaehyung Seo, Sugyeong Eo, and Heuiseok Lim. 2023. What is the Resultful Data?: Empirical Study on the Adaptability of the Automatic Post-Editing Training Data. In *Proceedings of the Eighth Conference on Machine Translation (WMT23)*, Singapore.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

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.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. 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.

Antonio Toral. 2020. Reassessing claims of human parity and super-human performance in machine translation at WMT 2019. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 185–194, Lisboa, Portugal. European Association for Machine Translation.

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.

Jiawei Yu, Min Zhang, Yanqing Zhao, Xiaofeng Zhao, Yuang Li, Chang Su, Yinglu Li, Miaomiao Ma, Shimin Tao, and Hao Yang. 2023. HW-TSC's Participation in the WMT 2023 Automatic Post Editing Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT23)*, Singapore.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.