Together We Can: Multilingual Automatic Post-Editing for Low-Resource Languages

Sourabh Deoghare[1], **Diptesh Kanojia**[2] **and Pushpak Bhattacharyya**[1]

[1]CFILT, Indian Institute of Technology Bombay, Mumbai, India

[2]Institute for People-Centred AI, University of Surrey, United Kingdom

{sourabhdeoghare, pb}@cse.iitb.ac.in, d.kanojia@surrey.ac.uk

Abstract

This exploratory study investigates the potential of multilingual Automatic Post-Editing (APE) systems to enhance the quality of machine translations for low-resource Indo-Aryan languages. Focusing on two closely related language pairs, English-Marathi and English-Hindi, we exploit the linguistic similarities to develop a robust multilingual APE model. To facilitate cross-linguistic transfer, we generate synthetic Hindi-Marathi and Marathi-Hindi APE triplets. Additionally, we incorporate a Quality Estimation (QE)-APE multitask learning framework. While the experimental results underline the complementary nature of APE and QE, we also observe that QE-APE multitask learning facilitates effective domain adaptation. Our experiments demonstrate that the multilingual APE models outperform their corresponding English-Hindi and English-Marathi single-pair models by 2.5 and 2.39 TER points, respectively, with further notable improvements over the multilingual APE model observed through multi-task learning (+1.29 and +1.44 TER points), data augmentation (+0.53 and +0.45 TER points) and domain adaptation (+0.35 and +0.45 TER points). We release the synthetic data, code, and models accrued during this study publicly for further research[1].

1 Introduction

Automatic Post-Editing (APE) aims to address the limitations of an unknown machine translation (MT) system by automatically correcting recurrent translation errors, thereby enhancing the overall translation output (Bojar et al., 2015). WMT has been driving the research in this field through APE-shared tasks over the years (Bojar et al., 2016; Chatterjee et al., 2018a, 2020; Bhattacharyya et al., 2023). However, as the generation of authentic (human-annotated) APE data is difficult, publicly available APE data is limited in terms of language pairs and sizes. The unavailability of an adequate amount of APE data for closely related languages has resulted in research focusing on different APE models for different language pairs.

Therefore, while we see considerable research devoted to advancing Neural Machine Translation (NMT) systems through multilingual MT (Johnson et al., 2017; Aharoni et al., 2019; Dabre et al., 2020; Costa-jussà et al., 2022; Gala et al., 2023), to the best of our knowledge, multilingual APE (MAPE) has not studied yet. Akin to how using multiple languages together helps improve translation capabilities, extending this approach to the APE task should enhance the translation correction capabilities for closely related languages despite limited resources.

This study focuses on English-Hindi (**En-Hi**) and English-Marathi (**En-Mr**) APE. Hindi and Marathi are two prominent Indo-Aryan languages with notable linguistic similarities because of their common linguistic ancestry (Chatterji, 1969; Masica, 1993). For example, both languages utilize the Devanagari script for writing and exhibit substantial vocabulary overlap (Kanojia et al., 2020b,a), largely due to their shared roots in Sanskrit. Grammatically, they both follow the Subject-Object-Verb (SOV) sentence structure, providing a familiar syntactic framework for speakers of either language (Subbārāo, 2012).

Such linguistic similarities enable cross-linguistic transfer and suggest that advancements in Multilingual APE (**MAPE**) could be particularly beneficial for improving translation quality across both these language pairs. Our contributions are:

1. *A multilingual APE framework* with data augmentation technique for low-resource language pairs with closely related target languages. The MAPE model trained on English-Hindi and English-Marathi pairs outperforms

[1]Multilingual-APE

Findings of the Association for Computational Linguistics: EACL 2024, pages 10800–10812

their corresponding single-pair APE models by 3.03 and 2.84 TER points, respectively (Refer Table 1, Table 2).

2. An extension of the *multitask-learning-based APE-QE framework* to MAPE. A multilingual model trained on En-Hi and En-Mr APE and QE tasks surpasses the performance of the single-task-based MAPE model by 1.29 and 1.44 TER points for En-Hi and En-Mr pairs, respectively (Refer Table 3).

3. A multitask learning-based *domain-adaptation technique* to cater to resource-constrained settings. Fine-tuning the MAPE model in the multitask-learning fashion only for the domain of interest improves overall performance over the non-adapted MAPE model by 0.35 and 0.45 TER points (Refer Table 4).

2 Related Work

WMT has been driving the APE research over the years through the APE shared tasks (Akhbardeh et al., 2021; Bhattacharyya et al., 2022, 2023). We see a paradigm shift from the WMT18 APE shared task onwards, where supervised transformed-based APE approaches became prominent for the task of improving high-quality translations obtained from neural MT systems (Chatterjee et al., 2018a).

Most neural APE approaches leverage transfer learning through the use of pre-trained language or translation models in APE model training (Lopes et al., 2019; Wei et al., 2020; Sharma et al., 2021; Deoghare and Bhattacharyya, 2022). To obtain latent representations of source and target sentences, Lee et al. (2020) employ multilingual or cross-lingual models. While these approaches followed a two-step training approach where the models are initially trained on synthetic APE data and then on authentic APE data, Oh et al. (2021) demonstrated an increasingly complex multi-step training approach, referred to as the Curriculum Training Strategy (CTS), under which a model is gradually trained on more and more advanced tasks enhances APE performance. Yang et al. (2020); Deoghare and Bhattacharyya (2022) demonstrated that increasing APE data with external MT candidates or through phrase-level APE triplets boosts feature diversity. Additionally, Yang et al. (2020); Huang et al. (2022) showed that incorporating domain information into the decoder adapter layers aids in domain adaptation of APE outputs. While we see the use of multilingual pre-trained models, the aim has always been to develop a single-pair APE model.

We also see some work in the literature on the joint exploration of APE and QE. Martins et al. (2017) used APE for the QE improvement. Hokamp (2017) used an ensemble of factored NMT models for word-level QE and APE tasks. Chatterjee et al. (2018b); Deoghare et al. (2023) compared multiple QE-APE coupling techniques and explored the extent of the complimentary nature of these two tasks. Deoghare et al. (2023) proposed a QE-APE multitask learning framework to train a single model on QE and APE tasks jointly through a sophisticated multitask learning approach, Nash-MTL (Navon et al., 2022).

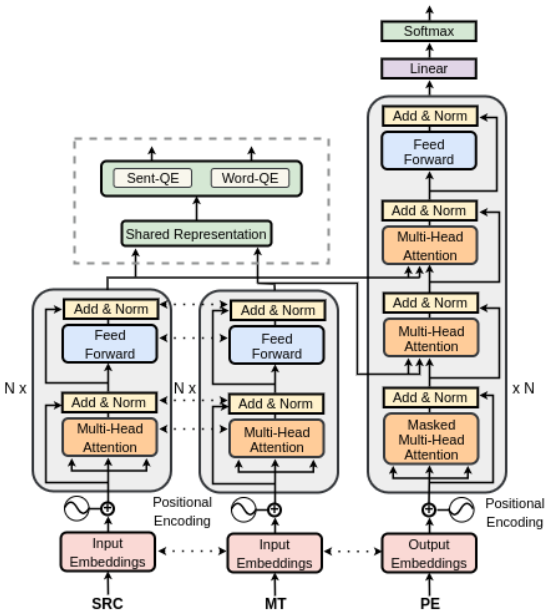

Figure 1: APE model architecture (Deoghare et al., 2023)

3 Methodology

Neural APE approaches model the APE problem as a multi-source translation task. For each pair of the input source sentence and its machine-generated translation as $x = \{x_i\}_{i=1}^{T_x}$, $z = \{z_i\}_{i=1}^{T_z}$, with lengths T_x, and T_z, respectively, the APE is trained in a supervised fashion to generate the post-edited translation $\hat{y} = \{\hat{y}_i\}_{i=1}^{T_{\hat{y}}}$ based on the reference post-edit $y = \{y_i\}_{i=1}^{T_y}$ of length T_y. For training the model, the cross-entropy loss is used as shown in Equation 1.

$$L_{APE} = -\sum_{w=1}^{|S|}\sum_{e=1}^{|V|} y_{w,e}\log\left(\hat{y}_{w,e}\right) \quad (1)$$

Where $|S|$ and $|V|$ denote the length of the sequence and the vocabulary size, respectively. $\hat{y}_{w,e}$, and $y_{w,e}$ represent the APE output and its reference, respectively.

While APE aims to generate corrected translations, QE systems are focused on assessing the *extent of correction* required (Sentence-level QE) and precise identification of *locations where those corrections are required* (Word-level QE). The task-specific head is added to the representations obtained from the final encoder layer to perform these tasks.

The Sentence-level QE task is modeled as a regression task with the goal of predicting the Direct Assessment (DA) score for given source sentences and their translation. Equation 2 shows the Mean Squared Error (MSE) loss used for this task.

$$\mathcal{L}_{sent} = MSE\Big(y_{da}, \hat{y}_{da}\Big) \quad (2)$$

Where $\hat{y}_{da}$ and y_{da} are the predicted DA score and the ground truth, respectively.

The Word-level Quality Estimation task is treated as a token-level classification task in which each token is classified into OK/BAD tags to denote its correctness. The cross-entropy loss utilized for this task is shown in Equation 3.

$$\mathcal{L}_{word} = -\sum_{i=1}^{2}\Big(y_{word}\odot\log(\hat{y}_{word})\Big)[i] \quad (3)$$

Where y_{word} and $\hat{y}_{word}$ denote ground truths and predicted tags.

This complementary nature of APE and QE tasks has been exploited in the QE-APE multi-task learning-based framework proposed by Deoghare et al. (2023) in which a two-encoder single-decoder model is jointly trained on APE, sentence-level QE, and word-level QE tasks. They observed higher APE performance when a sophisticated multitask learning algorithm, referred to as Nash-MTL (Navon et al., 2022), is used instead of linear scalarization where the task-specific losses are simply added together as shown in Equation 4. We use the Nash-MTL approach as described in Deoghare et al. (2023).

$$L_{LS-MTL} = L_{sent} + L_{word} + L_{APE} \quad (4)$$

Model Architecture In our work, we follow the approach proposed by Deoghare et al. (2023) for training the APE models and also use the same model architecture as shown in Figure 1. The architecture without the components shown in the dashed rectangle illustrates the architecture of the non-multitask learning-based APE model before the fine-tuning stage. When fine-tuning, no changes are made to the architecture in the case of single-task experiments. However, the QE task-specific heads are added to the shared representation layer that receives inputs from the final encoder layers for the multitask learning-based experiments.

Except for the transfer learning-based experiments, the encoder weights are initialized using IndicBERT (Kakwani et al., 2020), and the decoder weights are initialized randomly. We train all models using the Curriculum Training Strategy (CTS), same as Deoghare et al. (2023).

We consider the following three **baselines** in our work.

Do Nothing This baseline treats the original machine translations passed to the model as APE outputs.

Baseline APE In this experiment, we train separate APE models for English-Hindi and English-Marathi pairs. We consider these APE systems as **primary baselines**.

Transfer Learning To train an APE model for one language pair, we use the trained APE checkpoint of the other language pair. For example, to train an En-Mr APE model, we use the *En-Hi Baseline APE* and vice versa.

4 Multilingual Automatic Post-Editing

In the hope of getting help from the shared linguistic features of Hindi and Marathi, we perform various MAPE experiments hierarchically.

We leverage synthetic and authentic APE triplets of both En-Hi and En-Mr language pairs for all MAPE experiments. We combine and shuffle the synthetic APE datasets of both language pairs to create ***multilingual synthetic APE dataset***. Similarly, we combine and shuffle the authentic APE datasets of both pairs to create ***multilingual authentic APE dataset***. These combined sets of multilingual synthetic and authentic APE datasets are then used to train a MAPE model.

As an **initial approach**, we do not provide the model any indication about the language ID on the target side and represent it as **w/o-LangID** in further discussion.

Extending Authentic Triplets w/-LangID (Only Authentic w/-LangID) This experiment uses multilingual synthetic APE data as used in the *w/o-LangID* experiment. However, we prepend a respective language ID to a source sentence in each multilingual authentic APE dataset triplet.

Extending Synthetic and Authentic Triplets w/-LangID (w/-LangID) In this experiment, we prepend a respective language ID to a source sentence in each APE triplet of the multilingual synthetic and multilingual authentic APE dataset.

Based on the experimental results, we perform further MAPE exploration considering **w/-LangID** as the base experiment.

4.1 Data Augmentation

Through the following experiment, we try to assess the effectiveness of our proposed data augmentation approaches.

For the data augmentation, we translate a randomly picked subset of 0.5M (based on empirical findings; Refer Appendix A) source sentences from the synthetic APE triplets of each pair using IndicTrans2 (Gala et al., 2023) NMT system into the 'cross-target-language.' We refer to these translations as *External translations*. For example, we translate the source sentences from the English-Hindi synthetic APE data into the Marathi language to get quadruples of the following form: <*English source, Marathi External translation, Hindi translation, Hindi post-edit*>. Similarly, we translate the source sentences from the English-Marathi synthetic APE data into Hindi to get the <*English source, Hindi External translation, Marathi translation, Marathi post-edit*> quadruples.

The data augmentation is performed only during the second stage of CTS, in which the model is trained using the multilingual synthetic APE data. The authentic APE data is not augmented.

w/-LangID + Additional Pairs In this experiment, we exploit the cross-lingual information to improve the understanding of the source encoder by providing the APE model triplets of additional language pairs.

We use the quadruples to form Hindi-Marathi (<*Hindi External translation, Marathi translation, Marathi post-edit*>) and Marathi-Hindi (<*Marathi external translation, Hindi translation, Hindi post-edit*>) APE triplets. These triplets are added to the combined English-Hindi and English-Marathi synthetic APE datasets, and the entire set is shuffled before being used for the training. As in the case of the *w/-LangID* experiment, the source sentences in the augmented triplets are also prefixed with their respective language ID tag.

Therefore, during the second stage of the CTS, the model is trained on English-Hindi, English-Marathi, Hindi-Marathi, and Marathi-Hindi APE data.

w/-LangID + External Candidates We extend the data augmentation approach proposed by Yang et al. (2020) to the MAPE setting. In this experiment, we try to enhance the quality of the translation encoder by additionally providing it with the cross-lingual translation by appending it to the existing translation after adding a '<*sep*>' token. Therefore, the synthetic English-Hindi APE triplets are modified as follows: <*English source, Hindi translation* <*sep*> *Marathi external translation, Hindi post-edit*>. Similarly, the synthetic English-Marathi APE triplets take the following form: <*English source, Marathi translation* <*sep*> *Hindi external translation, Marathi post-edit*>.

4.2 APE-QE Multitask Learning

The sentence-level and word-level QE tasks have been shown to be helpful in improving the performance of APE systems in a scenario where the model is jointly trained for the QE and APE tasks in a single-pair setting. Through this experiment, we explore whether using sentence-level and word-level QE tasks along with APE also leads to improvements in the MAPE scenario. The following two experiments add to the setting of the *w/-LangID* experiment.

These experiments add multitask-learning to the *w/-LangID* experiment. During the fine-tuning stage of the CTS, the MAPE model is trained on the sentence-level QE and word-level QE tasks as well. We set the sentence-level QE loss to zero for those instances in the authentic APE dataset for which the Direct Assessment annotations are unavailable.

We perform the following two experiments under the multitask learning setting. (1) **MTL-MAPE (LS-MTL)**: The linear scalarization is used for combining task-specific losses. (2) **MTL-MAPE**

(Nash): Nash-MTL technique is used to combine the gradients and update the model weights. Refer to Navon et al. (2022) or Deoghare et al. (2023) for details on Nash-MTL.

4.3 Domain Adaptation

In this experiment, which we refer to as **DomainAdapt**, we explore the possibility of improving the domain-specific performance of MAPE through the domain-specific sentence-level and word-level QE information incorporated through multitask learning. In this experiment, during the fine-tuning stage of the CTS, the model is trained only on the domain-specific APE-QE instances of both language pairs. Similar to the *MTL-MAPE* experiments, we set the sentence-level QE loss to zero for those instances in the authentic APE dataset for which the Direct Assessment annotations are unavailable. Therefore, through this experiment, we get a separate APE model for each domain.

Since we end up with only a few training instances for each domain during fine-tuning, updating all network parameters during fine-tuning results in over-fitting. To avoid it, we add a single adapter layer in each decoder block as in Huang et al. (2022) and only update these adapter layers during fine-tuning. The overhead in performing this experiment is we get a separate domain-specific checkpoint. We combine news- and tourism-domain authentic APE triplets for this experiment for both pairs. Also, we combine the En-Mr health and En-Hi law and tourism domain triplets to form a 'General' domain authentic MAPE data.

5 Datasets

The MT and APE datasets of English-Hindi and English-Marathi are smaller than other language pairs like English-German, English-Russian, and English-Chinese. Considering the amount of MT parallel corpus, synthetic APE triplets, and authentic APE triplets for these language pairs, we consider English-Hindi and English-Marathi to be low-resource language pairs.

For English-Marathi APE, we use datasets released through the WMT22 (Bhattacharyya et al., 2022) English-Marathi APE shared task. The APE data consists of 18K authentic (human-generated post-edits) and 2.5M synthetic (Negri et al., 2018) APE triplets. The authentic APE triplets come from the News (6.5K), Tourism (6.5K), and Health (5K) domains, and the synthetic APE data triplets from multiple domains.

We use an in-house created APE corpus consisting of 9K training triplets for the English-Hindi pair, as an open-source APE corpus for this language pair is unavailable. Source sentences used in the authentic APE corpus are taken from publicly available data. 2.25K authentic triplets come from each of the News, Tourism, Law, and Education domains. Also, the MT system used for translating the source sentences is open-source. Apart from the training set, additional 1K-1K authentic triplets are set aside to be used as development and test sets. The source sentences in these datasets cover multiple domains. Additionally, we generate multi-domain 2.5M synthetic APE triplets from a subset of the Anuvaad[2] parallel corpus following the approach proposed by Negri et al. (2018). Along with this corpora, its details, annotation process, and guidelines are planned to be released through the upcoming WMT APE shared task. Further, we also use the BPCC (Gala et al., 2023) corpus for both pairs to train an MT model during the first stage of the CTS.

We use the English-Marathi and English-Hindi sentence-level QE data released through the WMT22 Zerva et al. (2022) and WMT23 (Blain et al., 2023) QE DA shared tasks, respectively, to get the DA scores for the APE triplets. We generate the word-level QE tags for both the pairs through the QE-corpus-builder tool[3].

6 Results and Discussion

This section discusses the results of different experiments. For the quantitative evaluation, we consider TER (Snover et al., 2006) as the primary metric. We also report BLEU (Papineni et al., 2002) scores. We perform a statistical significance test considering primary metric (TER) using William's significance test (Graham, 2015). The model training approach and the training details used are discussed in Appendix A.

Multilingual APE Table 1 compiles the results of initial multilingual experiments performed to investigate the impact of multilingual training and the importance of embedding explicit information about the target language into the MAPE model input. The model weight initialization through the

[2] Anuvaad Parallel Corpus

[3] https://github.com/deep-spin/qe-corpus-builder

Technique	En-Hi		En-Mr	
	TER	BLEU	TER	BLEU
Do Nothing	27.76	58.86	22.93	64.51
Baseline APE	20.85	66.70	20.58	66.95
Transfer Learning	20.59	66.99	20.25*	67.38
w/o-LangID	19.46	68.18	19.55	68.11
Only Authentic w/-LangID	18.93	68.91	18.73	69.04
w/-LangID	**18.35**	**69.56**	**18.19**	**69.56**

Table 1: Results of English-Hindi and English-Marathi APE systems on their respective evaluation sets. * denotes that the improvement over the primary baseline (*Baseline APE*) is insignificant (p being 0.05)

	En-Hi		En-Mr	
	TER	BLEU	TER	BLEU
Do Nothing	27.76	58.86	22.93	64.51
Baseline APE	20.85	66.70	20.58	66.95
w/-LangID	18.35	69.56	18.19	69.56
w/-LangID + Additional Pairs	**17.82**	**70.10**	**17.74**	**70.23**
w/-LangID + External Candidates	17.94	69.82	17.98*	69.77

Table 2: Results of English-Hindi and English-Marathi APE systems on their respective evaluation sets. * denotes the improvement over the *w/ LangID* is insignificant (p being 0.05).

other trained APE model (*Transfer Learning*) does not yield robust improvements even in the case of close target languages. The improvements are significant in the case of the En-Hi APE, while they are insignificant in the case of the En-Mr APE. It hints at the En-Mr APE model or data being more helpful to En-Hi APE than the other way around. A possible reason for such an outcome could be comparatively less authentic training data for En-Hi APE and a relatively easy evaluation set.

Comparison between the single-pair APE models and the MAPE model trained without inputting explicit information about the target language of the triplets reveals the potential of multilingual training in the case of closely related target languages. For both pairs, we observe more than 1 TER point improvement. In this comparison, we also observe En-Hi APE achieving more improvements than En-Mr APE.

The positive impact of presenting explicit information about the target language is evident from the results. Greater improvements when this information is presented to the model for all the triplets (synthetic and authentic) over the improvements when passed only for the authentic APE triplets highlight the need for target language token prefixing to the APE model while training in a multilingual setup.

The MAPE results show impressive improvements of 2.5 and 2.39 TER points over their corresponding single-pair APE models for En-hi and En-Mr, respectively, underlining the effectiveness of multilingualism in the case of linguistically similar languages.

Data Augmentation Results of data augmentation experiments are depicted in Table 2. The results show better performance gains through cross-lingual transfer when the multilingual model is trained on additional post-editing directions. Though not great, we observe modest gains of around 0.5 TER points for both pairs. Unlike the vanilla MAPE experiments, in the case of these experiments, we observe both pairs benefiting equally.

Comparatively, data augmentation through external candidates does not seem to be of much help as the model shows insignificant improvement in performance for En-Mr and small gains for the En-Hi pair. A possible reason could be that the translation encoder is already exposed to a vast amount of Hindi and Marathi text.

Multitask Learning Table 3 shows the extent of benefit that APE receives in the multilingual scenario from QE tasks through multitask learning. Similar to how joint training on QE and APE tasks has been shown to benefit APE in the case of single-pair APE models, results show its usefulness in the MAPE setting as well. Even the MAPE model trained through the simplest multitask learning approach, where the task-specific losses are added together, shows significant improvements over the non-multitask learning-based MAPE model. As expected, we observe further performance improvements for both pairs when a more sophisticated approach is employed to perform multitask learning.

When we use the MAPE model trained using the data augmentation through additional translation directions and fine-tune it through multitask learning, we do not see significant improvements over the multitask learning-based MAPE model trained without the data augmentation. This could be due to the law of diminishing returns. However, through this combination, we achieve improvements of more than 3 TER points for both pairs over their single-pair-based APE counterparts.

To the best of our knowledge, 16.75 TER points on the WMT22 English-Marathi APE development set are the highest performance achieved so far.

Technique	En-Hi		En-Mr	
	TER	BLEU	TER	BLEU
Do Nothing	27.76	58.86	22.93	64.51
Baseline APE	20.85	66.70	20.58	66.95
w/-LangID	18.35	69.56	18.19	69.56
MTL-MAPE (LS-MTL))	17.60	70.37	17.36	70.52
MTL-MAPE (Nash)	17.28	70.75	16.90	71.01
MTL-MAPE (Nash) + DataAug	**17.06**	**70.96**	**16.75**	**71.14**

Table 3: Results of Multitask Learning-based multilingual En-Hi and En-Mr APE experiments. The last row refers to an additional experiment where Nash-MTL-based multitask learning is used during the fine-tuning stage in the *w/-LangID + Additional Pairs* experiment. All improvements are significant with respect to the *w/-LangID* experiment (p being 0.05).

Technique	En-Hi		En-Mr	
	TER	BLEU	TER	BLEU
Do Nothing	27.76	58.86	22.93	64.51
Baseline APE	20.85	66.70	20.58	66.95
w/ LangID	18.35	69.56	18.19	69.56
DomainAdapt	**18.00**	**70.01**	**17.74**	**69.98**

Table 4: Results of Multitask Learning-based domain adaptation multilingual En-Hi and En-Mr APE experiments. All improvements with respect to w/ LangID are significant (with p being 0.05).

Domain Adaptation Table 4 shows the result of the domain adaptation experiment. The TER and BLEU scores reported in the table in the *DomainAdapt* row are an average of domain-specific scores obtained from separate domain-specific checkpoints. The domain adaption through multitask learning yields significant but modest performance improvements. Even though the adapters are used during the fine-tuning to prevent overfitting, they do not surpass or achieve comparable performance as we observe for a model, which is trained through data augmentation and fine-tuned through the QE-APE multitask learning.

Case Analysis Figure 2 shows two representative English-Hindi and English-Marathi APE examples to see whether the MAPE TER improvements reflect in the model outputs and also to investigate how multilingualism and multilingual QE-APE joint training helps the MAPE model.

The Hindi machine translation of the English source sentence is literal and misses the sense of the entity's departure being forever. The *Baseline APE* does not make any changes to the translation, which shows it failed to recognize the quality of the translation. Impressively, the MAPE successfully understands the literal translation of 'good' is incorrect and correctly uses the Hindi word *'sadaa'* instead of *'achche'* in the post-edit. It suggests how multilingual training improves the overall understanding of the target languages.

However, we see the MAPE model deviating from the *principle of minimal editing* as it unnecessarily changes the Hindi translation of the word 'departure' from *'jaanaa'* to its synonym *'prasthaan'*. It suggests the need to take steps to mitigate the problem of over-correction. While *'prasthaan'* word is present in both languages with the same meaning, its usage in Marathi is relatively more. The presence of the word *'prasthaan'* in the English-Marathi *Baseline APE* output again points to the knowledge sharing from English-Marathi to English-Hindi, due to multilingual training.

The MTL-MAPE model output where the translation of 'departure' is not touched and the correct translation to convey the sense of 'good' shows that training the MAPE model jointly on APE and QE tasks helps the model to follow the 'principle of minimal editing.' While the reference post-edit is far from the *MTL-MAPE* output, both sentences can be considered good translations.

On the English-Marathi side, the machine translation of the same source sentence is literal and also not fluent. The English-Marathi *Baseline APE* corrected all the fluency issues in the translation. However, it still is the literal translation of the source. The *MAPE* model takes care of the literal translation issue by making the necessary changes. As observed in the case of English-Hindi output of the *MAPE*, we observe the use of *'prasthaan'* instead of using the same Marathi phrase *'nighun jaane'* in the Marathi post-edit as well. Further, we see this issue being taken care of by the *MTL-MAPE* model, as also seen in the case of the Hindi post-editing.

Such similarity in the behavior of the *MAPE* and *MTL-MAPE* models suggest that the MAPE training might be helping the models in generating consistent post-edits across the target languages when the source sentence is the same, and the quality of the machine translations is similar. While not reported in the paper, we would like to share that our analysis of a small subset of instances and their corresponding MAPE outputs strengthens this conjecture. However, a large-scale scientific analysis is required to confirm this claim. We provide additional examples in Appendix C.

English-Hindi Example			
Source	This time his departure was for good.		
Machine Translation	इस बार उनका जाना अच्छे के लिए था।	Iss *(This)* baar *(time)* unakaa *(his)* jaanaa *(departure)* achche *(good)* ke *(of)* liye *(for)* thaa *(was)* ।	This time his departure was for good.
Reference	इस बार उनकी विदाई हमेशा के लिए हुई।	Iss *(This)* baar *(time)* unaki *(his)* widaayi *(departure)* hameshaa *(forever)* ke *(of)* liye *(for)* hui *(happened)* ।	This time his departure was forever.
Baseline APE	इस बार उनका जाना अच्छे के लिए था।	Iss *(This)* baar *(time)* unakaa *(his)* jaanaa *(departure)* achche *(good)* ke *(of)* liye *(for)* thaa *(was)* ।	This time his departure was for good.
w/-LangID	इस बार उनका प्रस्थान सदा के लिए था।	Iss *(This)* baar *(time)* unakaa *(his)* prasthaan *(departure)* sadaa *(forever)* ke *(of)* liye *(for)* thaa (was) ।	This time his departure was forever.
MTL-MAPE (Nash-MTL)	इस बार उनका जाना सदा के लिए था।	Iss *(This)* baar *(time)* unakaa *(his)* jaana *(departure)* sadaa *(forever)* ke *(of)* liye *(for)* thaa (was) ।	This time his departure was forever.
English-Marathi Example			
Source	This time his departure was for good.		
Machine Translation	या वेळी त्यांनी निघून जाणे चांगली होती.	Yaa *(this)* veli *(time)* tyaani *(by him)* nighun *(to start leaving)* jaane *(to go)* chaangali *(good)* hoti *(was)* .	This time him was going was good.
Reference	या वेळी त्यांचे निघून जाणे कायमचे होते.	Yaa *(this)* veli *(time)* tyaanche *(his)* nighun *(to start leaving)* jaane *(to go)* kaayamache (of forever) hote *(was)* .	This time his departure was forever.
Baseline APE	या वेळी त्यांचे प्रस्थान चांगले होते.	Yaa *(this)* veli *(time)* tyaanche *(his)* prasthaan *(departure)* chaangale *(to go)* hote *(was)* .	This time his departure was good.
w/-LangID	या वेळी त्यांचे प्रस्थान कायमचे होते.	Yaa *(this)* veli *(time)* tyaanche *(his)* prasthaan *(departure)* kaayamache *(of forever)* hote *(was)* .	This time his departure was forever.
MTL-MAPE (Nash-MTL)	या वेळी त्यांनी निघून जाणे कायमसाठीचे होते.	Yaa *(this)* veli *(time)* tyaani nighun *(to start leaving)* jaane *(to go)* kaayamasaathiche *(of for forever)* hote *(was)* .	This time his departure was forever.

Figure 2: Comparison of English-Marathi and English-Hindi APE outputs obtained from Baseline APE and two MAPE systems.

7 Conclusion and Future Work

In this work, we explore multilingual APE, which has not yet been attempted to the best of our knowledge. We focus on two low-resource English-Indic language pairs where the target languages, Hindi and Marathi, are closely related Indo-Aryan languages. The results of our initial experiments show the effectiveness of multilingual training for improving APE performance of both pairs. Relatively higher improvements for English-Hindi than English-Marathi suggest that the MAPE can benefit the lower resource language pair in the APE setting, to improve upon their APE performance.

We observe further performance gains using the proposed simple yet effective data augmentation approach in which we generate synthetic Hindi-Marathi and Marathi-Hindi APE triplets, to enable cross-lingual transfer. Our investigation into the use of the QE-APE multitask learning framework in the multilingual APE setting reveals that the QE helps MAPE improve its capability to assess translation quality, and mitigate the problem of over-correction. The exploration of using only the domain-specific QE information to perform domain adaptation shows that modest performance gains can also be obtained, even in resource-constrained settings.

With 17.06 and 16.75 TER on English-Hindi and English-Marathi evaluation sets, our multilingual APE model achieves state-of-the-art performance for the English-Marathi pair, and establishes a strong baseline for English-Hindi. Also, our qualitative analysis hints at the possible future work-Using MAPE for consistent post-edits across languages.

The investigation performed in this work suggests that multilingual APE is a promising research direction for the advancement of APE systems for low-resource language pairs. Even though the gains of more than 3 TER points highlight the robustness of MAPE, in the future, this work can be extended to multiple low-resource language pairs to analyze the generalizability. An in-depth study could be conducted to analyze the extent of consistency of the MAPE model outputs across target languages when fed with the same source and similar quality translations. Further experiments could be conducted to gauge the optimal number of augmented triplets of specific additional language pairs to improve the MAPE performance on specific or all language pairs.

8 Limitations

The multilingual APE investigation undertaken under this work is limited to a specific case of two low-resource language pairs where the source language is English, and the target language is either Hindi or Marathi, which are linguistically closely related In-

dian languages. Due to the lack of APE resources for other language pairs, answering whether the simple multilingual APE training, as explored in this work, will result in impressive performance gains for other low-resource language pairs as well needs further exploration. Also, the quality of Baseline APE models of both pairs (20.85 and 20.58 TER points) is similar. Current work does not explore the improvements MAPE can bring when the *Baseline APE* models of each pair have different qualities.

9 Ethics Statement

Our APE models are trained on either in-house datasets or on the publicly available datasets referenced in this paper. These datasets have been previously collected and annotated; no new data collection has been carried out as part of this work. The in-house created dataset will be made public through the upcoming WMT APE shared task. Furthermore, these are standard benchmarks released in recent WMT shared tasks. No user information was present in any of the datasets used in the work, protecting the privacy and identity of users. Also, the synthetic data, code, and models generated as a part of this work will be released publicly under the CC-BY-SA 4.0 license for further research. We understand that every dataset is subject to intrinsic bias and that computational models will inevitably learn biased information from any dataset.

References

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

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.

Pushpak Bhattacharyya, Rajen Chatterjee, Markus Freitag, Diptesh Kanojia, Matteo Negri, and Marco Turchi. 2023. Findings of the WMT 2023 shared task on automatic post-editing. In *Proceedings of the Eighth Conference on Machine Translation*, pages 672–681, Singapore. Association for Computational Linguistics.

Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. 2023. Findings of the WMT 2023 shared task on quality estimation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, Singapore. 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, Aurélie Névéol, 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: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

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. 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 Track)*, pages 26–38, Boston, MA. Association for Machine Translation in the Americas.

S.K. Chatterji. 1969. *Indo-Aryan & Hindi*. Firma K. L. Mukhopadhyay.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Computing Surveys (CSUR)*, 53(5):1–38.

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 (WMT)*, pages 682–688, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Sourabh Deoghare, Diptesh Kanojia, Fred Blain, Tharindu Ranasinghe, and Pushpak Bhattacharyya. 2023. Quality estimation-assisted automatic post-editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1686–1698, Singapore. Association for Computational Linguistics.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Yvette Graham. 2015. Improving evaluation of machine translation quality estimation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1804–1813, Beijing, China. Association for Computational Linguistics.

Chris Hokamp. 2017. Ensembling factored neural machine translation models for automatic post-editing and quality estimation. In *Proceedings of the Second Conference on Machine Translation*, pages 647–654, Copenhagen, Denmark. Association for Computational Linguistics.

Xiaoying Huang, Xingrui Lou, Fan Zhang, and Tu Mei. 2022. LUL's WMT22 automatic post-editing shared task submission. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 689–693, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Diptesh Kanojia, Raj Dabre, Shubham Dewangan, Pushpak Bhattacharyya, Gholamreza Haffari, and Malhar Kulkarni. 2020a. Harnessing cross-lingual features to improve cognate detection for low-resource languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1384–1395, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Diptesh Kanojia, Malhar Kulkarni, Pushpak Bhattacharyya, and Gholamreza Haffari. 2020b. Challenge dataset of cognates and false friend pairs from Indian languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3096–3102, Marseille, France. European Language Resources Association.

Jihyung Lee, WonKee Lee, Jaehun Shin, Baikjin Jung, Young-Kil Kim, and Jong-Hyeok Lee. 2020. 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*, pages 777–782, Online. Association for Computational Linguistics.

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.

André F. T. Martins, Marcin Junczys-Dowmunt, Fabio N. Kepler, Ramón Astudillo, Chris Hokamp, and Roman Grundkiewicz. 2017. Pushing the limits of translation quality estimation. *Transactions of the Association for Computational Linguistics*, 5:205–218.

Colin P Masica. 1993. *The indo-aryan languages*. Cambridge University Press.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. 2022. Multi-task learning as a bargaining game. In *International Conference on Machine Learning*.

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).

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.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Abhishek Sharma, Prabhakar Gupta, and Anil Nelakanti. 2021. Adapting neural machine translation for automatic post-editing. In *Proceedings of the Sixth Conference on Machine Translation*, pages 315–319, Online. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Kārumūri V Subbārāo. 2012. *South Asian languages: A syntactic typology*. Cambridge University Press.

Daimeng Wei, Hengchao Shang, Zhanglin Wu, Zhengzhe Yu, Liangyou Li, Jiaxin Guo, Minghan Wang, Hao Yang, Lizhi Lei, Ying Qin, and Shiliang Sun. 2020. HW-TSC's participation in the WMT 2020 news translation shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 293–299, Online. Association for Computational Linguistics.

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*, pages 797–802, Online. Association for Computational Linguistics.

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.

No. of Triplets	En-Hi	En-Mr
0.1M	18.16	18.09
0.25M	17.95	17.78
0.5M	**17.82**	**17.74**
1M	17.89	17.81
2.5M	18.50	18.30

Table 5: TER scores of *w/-LangID + Additional Pairs* experiment on the respective evaluation set when different amounts of triplets are used for the data augmentation. An equal number of triplets Hindi-Marathi and Marathi-Hindi triplets are used in each experiment.

A Experimental Details

The experimental setup for conducting all APE experiments across English-Marathi (En-Mr) and English-Hindi (En-Hi) language pairs.

A.1 Training Approach

We utilize the Curriculum Training Strategy (CTS) as described by Deoghare et al. (2023) to train both APE and MAPE systems. The steps of CTS are outlined as follows:

Initially, we train a single-encoder, single-decoder model for the NMT task. We merge the parallel corpora of both language pairs to train a multilingual NMT. Each source sentence in the corpus is prefixed with the respective target Language ID ('hin_Deva' for Hindi, 'mar_Deva' for Marathi).

In the next step, we introduce the translation encoder and train the two-encoder single-decoder model for the APE task using synthetic APE data in two phases. In the first phase, the model is trained on a subset of the synthetic corpus consisting of triplets with poorer TER than the Do Nothing baseline. In the second phase, training continues on the remaining synthetic corpus subset, which includes triplets with equal or better TER than the Do Nothing baseline.

Finally, in the third step, we fine-tune the APE model using authentic APE data.

A.2 Training Details

We adhere to a standardized set of configurations to ensure uniformity in experimental settings across different experiments. The batch size is 32. We employ early stopping with a patience of 5 epochs and train the model for a maximum of 10000 epochs. The Adam optimizer is used with a learning rate of 5 x 10^{-5}, and β_1 and β_2 are to 0.9 and 0.997, respectively. We incorporate 15,000 warmup steps. For the beam search, the beam size is set to 5. The size of the adapter used in the domain adaptation experiments is 512. NVIDIA100 GPUs are utilized for experimentation purposes. The MAPE model consists of approximately 40 million parameters, and training with CTS takes around 56 hours. Given the extensive number of experiments and the significant time and resources required, we report single-run results.

For preprocessing the data, we utilize the NLTK library[4] for English text and the IndicNLP library[5] for Hindi and Marathi text. Model training and inference are carried out using Pytorch[6]. To compute TER scores, we use the official WMT APE evaluation script[7], and for BLEU scores, we use the SacreBLEU[8] library.

B Data Augmentation

Table B shows how the MAPE performance varies when the multilingual synthetic data is further augmented through different numbers of Hindi-Marathi and Marathi-Hindi synthetic APE triplets.

C Additional Examples

Figure 3 and 4 show English-Hindi and English-Marathi Baseline APE and MAPE outputs pointing out some of the limitations of the current MAPE system.

[4] https://www.nltk.org/

[5] https://github.com/anoopkunchukuttan/indic_nlp_library

[6] https://pytorch.org/

[7] https://github.com/sheffieldnlp/qe-eval-scripts

[8] https://github.com/mjpost/sacrebleu

English-Hindi Examples			
Source	The High Court passed an interim order restraining the defendant from transferring the disputed property.		
Machine Translation	उच्च न्यायालय ने प्रतिवादी को विवादित संपत्ति स्थानांतरित करने के लिए एक अंतरिम आदेश पारित किया।	Uchcha *(high)* nyaayaalay *(court)* ne *(by)* pratiwaadi *(defendant)* ko *(to)* weewadeet *(disputed)* sampattee *(property)* sthaanaantareet *(transfer)* karane *(to do)* ke *(of)* leeye *(for)* ek *(one)* antareem *(interim)* aadesh *(order)* paareet *(to pass)* keeya *(did)*.	The High Court passed an interim order to transfer the disputed property to the respondent.
Baseline APE	उच्च न्यायालय ने प्रतिवादी को विवादित संपत्ति स्थानांतरित करने से रोकने के लिए एक अंतरिम आदेश पारित किया।	Uchcha *(high)* nyaayaalay *(court)* ne *(by)* pratiwaadi *(defendant)* ko *(to)* weewadeet *(disputed)* sampattee *(property)* sthaanaantareet *(transfer)* karane *(to do)* se *(from)* rokane *(to stop)* ke *(of)* leeye *(for)* ek *(one)* antareem *(interim)* aadesh *(order)* paareet *(to pass)* keeya *(did)*.	The High Court passed an interim order restraining the defendant from transferring the disputed property.
MTL-MAPE (Nash-MTL)	उच्च न्यायालय ने प्रतिवादी को विवादित संपत्ति का स्थानांतरण करने से रोकने हेतु अंतरिम आदेश पारित किया था।	Uchcha *(high)* nyaayaalay *(court)* ne *(by)* pratiwaadi *(defendant)* ko *(to)* weewadeet *(disputed)* sampattee *(property)* ka *(of)* sthaanaantaran *(transfer)* karane *(to do)* se *(from)* rokane *(to stop)* hetu *(with motive)* antareem *(interim)* aadesh *(order)* paareet *(to pass)* keeya *(did)* thaa *(had done)*.	The High Court had passed an interim order restraining the defendant from transferring the disputed property.
Comment	Even though the tense of the sentence is correct (present perfect or simple past) in the translation, MAPE has changed it (past perfect).		
Source	He was surprised to learn about the ancient traditions still practiced in the temples of Kerala.		
Machine Translation	वह केरल के मंदिर के बारे में प्रचलित पुराने परंपराओं के बारे में जानकर आश्चर्यचकित रह गए थे।	Wah *(he)* keral *(kerala)* ke *(of)* mandeer *(temple)* ke *(of)* baare *(about)* main *(in)* prachaleet *(practiced)* puraane *(old)* paramparaaon *(traditions)* ke *(of)* baare *(about)* main *(in)* jaanakar *(by knowing)* aashcharyachakeet *(surprised)* rah *(to be)* gaye *(went)* the *(were)*.	He was surprised to know about the old traditions prevalent in the temple in Kerala.
Baseline APE	वह केरल के मंदिर में अब भी प्रचलित प्राचीन परंपराओं के बारे में जानकर आश्चर्यचकित थे।	Wah *(he)* keral *(kerala)* ke *(of)* mandeer *(temple)* main *(in)* ab *(now)* bhi *(still)* prachaleet *(practiced)* praachin *(ancient)* paramparaaon *(traditions)* ke *(of)* baare *(about)* main *(in)* jaanakar *(by knowing)* aashcharyachakeet *(surprised)* the *(were)*.	He was surprised to learn about the ancient traditions still practiced in the temple of Kerala.
MTL-MAPE (Nash-MTL)	वह केरल के मंदिर में अब भी प्रचलित प्राचीन परंपराओं के बारे में जानकर आश्चर्यचकित थे।	Wah *(he)* keral *(kerala)* ke *(of)* mandeer *(temple)* main *(in)* ab *(now)* bhi *(still)* prachaleet *(practiced)* praachin *(ancient)* paramparaaon *(traditions)* ke *(of)* baare *(about)* main *(in)* jaanakar *(by knowing)* aashcharyachakeet *(surprised)* the *(were)*.	He was surprised to learn about the ancient traditions still practiced in the temple of Kerala.
Comment	Like Baseline APE, MAPE failed to identify that the translation of the plural word 'temples' is in the singular form ('*mandeer*').		

Figure 3: Comparison of English-Marathi and English-Hindi APE outputs obtained from Baseline APE and MAPE systems.

English-Marathi Examples			
Source	The museum displayed a beautiful terracotta sculpture depicting an ancient warrior.		
Machine Translation	संग्रहालयात एक प्राचीन योद्धा दर्शविणारी सुंदर टेराकोटा शिल्पकला आहे.	Sangrahaalayaat *(in museum)* ek *(one)* praachin *(ancient)* yoddhaa *(warrior)* darshaveenaari *(depicting)* sundar *(beautiful)* teracota *(terracotta)* sheelpakalaa *(sculptural art)* aahe *(is)*.	The museum displays a beautiful terracotta sculptural art depicting an ancient warrior.
Baseline APE	संग्रहालयात प्राचीन योद्ध्याचे दर्शन घडवणारी सुंदर टेराकोटाची शिल्पकला दर्शविली होती.	Sangrahaalayaat *(in museum)* praachin *(ancient)* yoddhyaache *(of warrior)* darshan *(view)* ghadawanaari *(that does)* sundar *(beautiful)* teracotachi *(of terracotta)* sheelpakalaa *(sculptural art)* darshaweeli *(displayed)* hoti *(was)*.	The museum had displayed a beautiful terracotta sculptural art depicting an ancient warrior.
MTL-MAPE (Nash-MTL)	संग्रहालयात प्राचीन योद्ध्याचे दर्शन घडवणारी सुंदर टेराकोटाची शिल्पकला प्रदर्शित केली आहे.	Sangrahaalayaat *(in museum)* praachin *(ancient)* yoddhyaache *(of warrior)* darshan *(view)* ghadawanaari *(that does)* sundar *(beautiful)* teracotachi *(of terracotta)* sheelpakalaa *(sculptural art)* pradarsheet *(depicted)* keli *(has done)* aahe *(is)*.	The museum displays a beautiful terracotta sculpture depicting an ancient warrior.
Comment	MAPE output shows an adverse effect of multilingualism as it uses the Hindi-influenced word '*pradarshit*' as opposed to 'darshaweeleli' used by the Baseline.		
Source	Regular physical exercise is essential for maintaining overall health and preventing chronic diseases.		
Machine Translation	एकंदरीत आरोग्य राखण्यासाठी आणि जुनाट आजार रोखण्यासाठी नियमित शारीरिक व्यायाम करणे आवश्यक आहे.	Ekandarit *(overall)* aarogya *(health)* raakhanyaasaathi *(for maintaining)* aanee *(and)* junaat *(chronic)* aajaar *(diseases)* rokhanyaasaathi *(for stopping)* neeyameet *(regularly)* sharireek *(physical)* vyaayaam *(excercise)* karane *(to do)* aavashyak *(essential)* aahe *(is)*.	Regular physical exercise is essential for maintaining overall health and preventing chronic diseases.
Baseline APE	सर्वांगीण आरोग्य राखण्यासाठी आणि जुनाट आजार रोखण्यासाठी नियमित शारीरिक व्यायाम करणे आवश्यक आहे.	Sarwaangin (overall) aarogya (health) raakhanyaasaathi (for maintaining) aanee (and) junaat (chronic) aajaar (diseases) rokhanyaasaathi (for stopping) neeyameet (regularly) sharireek (physical) vyaayaam (excercise) karane (to do) aavashyak (essential) aahe (is).	Regular physical exercise is essential for maintaining overall health and preventing chronic diseases.
MTL-MAPE (Nash-MTL)	एकंदरीत आरोग्य राखण्यासाठी आणि जुनाट आजार रोखण्यासाठी नियमित शारीरिक व्यायाम करणे आवश्यक आहे.	Ekandarit (overall) aarogya (health) raakhanyaasaathi (for maintaining) aanee (and) junaat (chronic) aajaar (diseases) rokhanyaasaathi (for stopping) neeyameet (regularly) sharireek (physical) vyaayaam (excercise) karane (to do) aavashyak (essential) aahe (is).	Regular physical exercise is essential for maintaining overall health and preventing chronic diseases.
Comment	The translation is correct, so APE is not supposed to make any edits. While Baseline APE follows it, MAPE attempts to improve fluency by substituting '*sarwaangin*' for '*ekandarit*.'		

Figure 4: Comparison of English-Marathi and English-Hindi APE outputs obtained from Baseline APE and MAPE systems.