Refer to the Reference: Reference-focused Synthetic Automatic Post-Editing Data Generation

Sourabh Deoghare, **Diptesh Kanojia** and **Pushpak Bhattacharyya**

CFILT, Indian Institute of Technology Bombay, Mumbai, India

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

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

Abstract

A prevalent approach to synthetic APE data generation uses source (*src*) sentences in a parallel corpus to obtain translations (*mt*) through an MT system and treats corresponding reference (*ref*) sentences as post-edits (*pe*). While effective, due to independence between *mt* and *pe*, these translations do not adequately reflect errors to be corrected by a human post-editor. Thus, we introduce a novel and simple yet effective reference-focused synthetic APE data generation technique that uses *ref* instead of *src* sentences to obtain corrupted translations (*mt_new*). The experimental results across English-German, English-Russian, English-Marathi, English-Hindi, and English-Tamil language pairs demonstrate the superior performance of APE systems trained using the newly generated synthetic data compared to those trained using existing synthetic data. Further, APE models trained using a balanced mix of existing and newly generated synthetic data achieve improvements of 0.37, 0.19, 1.01, 2.42, and 2.60 TER points, respectively. We will release the generated synthetic APE data.

1 Introduction

Automatic Post-Editing (APE) aims to reduce the human-post-editing effort by correcting recurrent errors in translations generated by a Machine Translation (MT) system. APE systems are especially useful in a black-box scenario where the underlying MT system is inaccessible (Chatterjee et al., 2020). Utilizing transformer-based encoder-decoder models in a supervised fashion for generating a post-edited translation, given the source sentence and its MT-generated translation as inputs, is the prevalent approach for developing APE systems (Chatterjee et al., 2019, 2020; Bhattacharyya et al., 2022). Developing such robust APE systems requires an adequate amount of high-quality (authentic) APE corpus consisting of triplets: a source sentence (*src*), its translation obtained from an MT system (*mt*), and the corresponding human-post-edited translation (*pe*). However, obtaining the human post-edits is expensive in terms of time and money.

The *de-facto* method used to alleviate this problem is to artificially generate APE triplets using a parallel corpus containing source (*src*) and reference (*ref*) sentence pairs. This approach translates *src* into *mt* by using an MT system, and the *ref* is treated as *pe* (Negri et al., 2018). While this approach has been effective in improving APE performance (Wang et al., 2020; Lee et al., 2021), *pe* sentences in this data may not be minimally post-edited versions of the corresponding *mt* sentences (Lee et al., 2021). That means error patterns present in *mt* may differ from the *mt* sentences of authentic APE data. Appendix A contains a representative example showing the same. With a **motivation** to mitigate this problem, we propose a reference (*ref*)-focused synthetic APE data generation method.

Our contributions are:

1. A novel synthetic APE data generation technique that uses paraphrased versions of reference sentences to generate translations via a round-trip approach (Refer Section 3).

2. Comprehensive validation and analysis of our technique on five (En-De, En-Ru, En-Mr, En-Hi, and En-Ta) APE systems utilizing the newly generated data improving upon APE systems trained using existing synthetic APE data (primary baselines) where our technique shows improvement by 0.37, 0.19, 1.01, 2.42, and 2.60 TER points, respectively (Refer Table 3).

3. Public release of synthetic APE corpora for En-De, En-Ru, En-Mr, En-Hi, and En-Ta containing 4, 7.7, 2.5, 2.5, and 2.5 million triplets, respectively[1].

[1] Github Repository

Proceedings of the 31st International Conference on Computational Linguistics, pages 5123–5135
January 19–24, 2025.

The learning objective of APE and the model architecture are discussed in Appendix B.

2 Related Work

There have been a few efforts to address the problem of limited authentic APE data availability for developing APE systems by generating the data synthetically. Junczys-Dowmunt and Grundkiewicz (2016) used target-side monolingual data to generate APE triplets. The approach treats the original target language sentences as *pe*, and *mt* sentences are obtained via round-trip translations performed using target-source and source-target phrase-based SMT systems (Zens et al., 2002). Intermediate translations obtained from the target-source MT system are considered as *src* sentences. Due to its way of construction, by maintaining a weak connection between *mt* and *pe*, this data approximates the nature of the authentic APE data. However, the quality of *src* sentences in this data is dependent on the quality of the target-source MT system. Freitag et al. (2022) used round-trip translation to generate artificial erroneous translations and trained a monolingual APE model.

To mitigate this drawback, eSCAPE (Negri et al., 2018) synthesized APE triplets from a parallel corpus. This approach treats the *src* and *ref* in a parallel corpus as the *src* and *pe* segments of APE triplets. To obtain *mt*, the *src* segments are translated through an MT system. This method has been employed as a data augmentation technique in prior APE research, demonstrating impressive performance (Wang et al., 2020; Lee et al., 2021). However, since the translated sentences are independent of *pe*, the *mt* sentences may not accurately reflect errors observed in translations of an authentic APE data (Lee et al., 2021). Tuan et al. (2021) adds to the technique proposed by Negri et al. (2018) and uses an MLM-based approach called MLM-Rewrites to generate the triplets.

Lee et al. (2020, 2022a) tried to address the limitation of eSCAPE by randomly injecting insertion, deletion, shifting, and substitution noise into selected candidate *ref* sentences of a parallel corpus to obtain noisy *mt*. The original *src* and *ref* were treated as *src* and *pe* in the APE triplets. Similarly, Lin et al. (2022) used an approach that injects noise in the *ref* sentences through insertion, selection, transposition, and repetition perturbations for generating synthetic data for the task of automatic post-editing of human-generated translations.

A work by Lee et al. (2021, 2022b) proposed a method similar to back-translation for generating translations in APE triplets. The approach involves training a model to produce *mt* with *src* and *pe* as inputs. While this approach can generate *mt* containing practical errors corrected by humans, it requires APE data that contains human-post-edits for training the synthetic APE data generation model, limiting its applicability.

An approach proposed by Moon et al. (2022a,b) focuses on noising scheme-based data generation for pairs with English on the target side. This approach eliminates the need for a translation model in data generation and can produce *mt* with errors reflective of those encountered in actual corrections. However, this approach requires target language-specific linguistic resources, which may be difficult to gather for low-resource language pairs.

Inspired by the existing approaches, this work proposes a generalizable synthetic APE data generation technique that does not rely on linguistic resources of a source or target languages and uses *ref* sentences from a parallel corpus to obtain its corrupted version *mt_new*, which is treated as *mt*.

3 Reference-focused Synthetic APE Data Generation

Unlike the prevalent synthetic APE data generation approach (Negri et al., 2018), referred hereinafter as *'existing'* synthetic APE corpus, which utilizes a *src* and generates *mt* which is independent of *pe* or *ref*, our approach focuses on *ref* and follows a noise injection pipeline to obtain *mt_new*. Figure 1 shows the two stages of the synthetic data generation pipeline.

Training NMT Models In the first stage, a parallel corpus is extracted by picking *src* and *pe* sequences from each triplet of the existing synthetic APE corpus. This extracted parallel corpus, along with optional additional parallel corpus, is used to train *src* to *ref* and *ref* to *src* Neural Machine Translation (NMT) systems (Vaswani et al., 2017). Use of the same *src* and *pe* sentences from the existing synthetic corpus for training both the NMT systems is done so that the translations generated from these NMT systems will be closer to the original *src* and *ref* sentences.

In the second stage, we generate erroneous translations by injecting noise into *pe*, i.e., *ref* sentences. The process is divided into two steps.

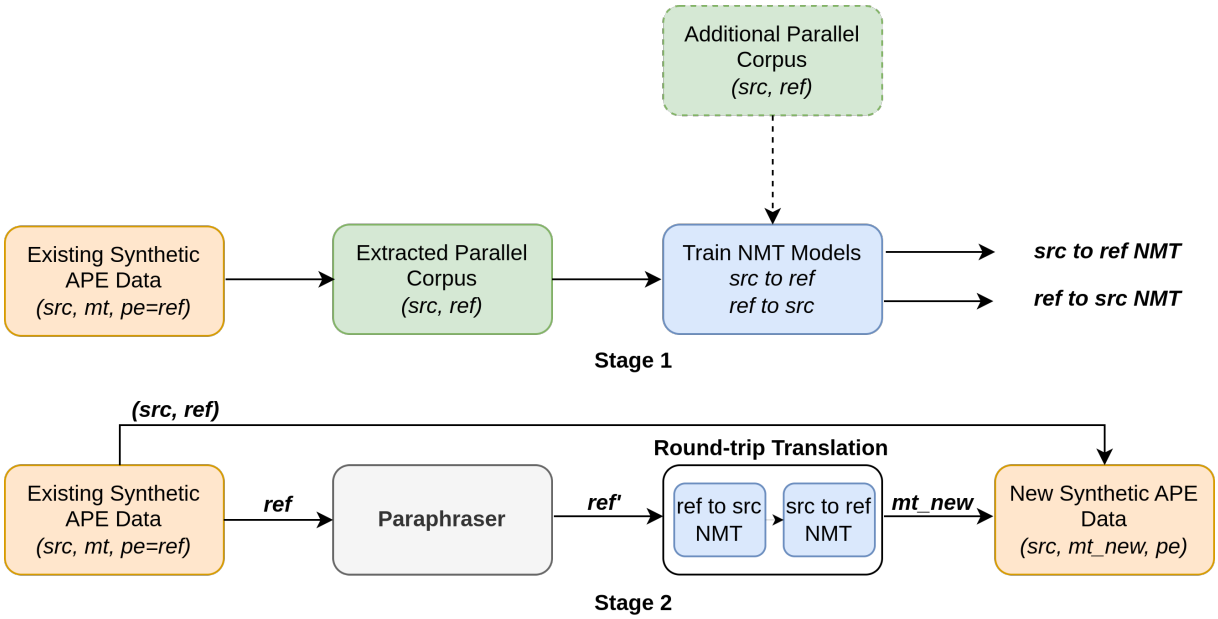

Figure 1: Proposed Synthetic APE data generation pipeline. Stage 1 consists of training *src-ref* and *ref-src* MT systems. The block shown via the dashed line denotes using an additional MT parallel corpus is optional. In Stage 2, *pe* sentences in the existing synthetic APE dataset are paraphrased, and then a round-trip translation is performed using the trained NMT systems to get the *mt_new*.

Paraphrasing In the first step, we pass *ref* to a neural paraphraser (Egonmwan and Chali, 2019). As the *ref* or *pe* sentences, using which corresponding *mt* sequences need to be generated, are used for training NMT systems in the first stage, directly injecting noise into original *pe* sentences could limit the feature diversity of the generated *mt_new*.

To avoid it, we use a paraphraser. Given an n-word sequence $S = (s_1, s_2, ..., s_n)$, the aim of the paraphraser is to generate another sequence $S' = (s'_1, s'_2, ..., s'_m)$ having m words that convey the same underlying meaning as S.

Thus, passing a *ref* to a paraphraser allows obtaining a semantically similar sequence (*ref'*) that differs from the original sequence in terms of **lexical choice, length, and word order**.

Round-trip Translation (RTT) The second step consists of passing the *ref'* through *ref-src* and *src-ref* MT systems trained during the first stage, respectively. The sequence obtained after this round-trip translation, *mt_new*, which is likely to be a corrupted version of the *pe* is considered as a machine-generated translation to the form of an APE triplet consisting of *(src, mt_new, pe)*.

Why do we need Paraphrasing? Since the RTT is performed using the target-to-source and source-to-target NMT systems, which are trained using the *src* and *ref* from the *existing* synthetic APE corpus only, directly passing the same *ref* sentences which have seen both NMT models during their training, results in generating same or very similar sequences as inputs *ref* sequences through the RTT. Also, even though the RTT-generated sequences may not exactly be the same as the input, the possibility of error coverage that such RTT-generated sequences could provide is limited. Therefore, it is necessary to obtain *ref'*, which differs from the original *ref*.

An approach like random masking of tokens in *ref* and then using a language model with MLM objective to fill the masked positions limit the variations that can be brought to the original *ref*. Such techniques do not introduce word order-based changes. Therefore, we opted for using a paraphrase to obtain *ref'*.

Further, passing *ref'*, which is likely to differ from *ref* in terms of lexical choice, length, and word order, to RTT increases the odds of obtaining *mt_new*.

Since the NMT model generates *mt_new*, unlike other approaches that induce noise into references through MLM-like approaches, it reflects errors we find in usual NMT outputs.

Corpus Interleaving We expect that the corrupted translations generated using the Paraphras-

	En-De	En-Ru	En-Mr	En-Hi	En-Ta
Synthetic	4M	7.7M	2.5M	2.5M	2.5M
Authentic	7K	15K	16K	7K	7K
Dev/Test	1K	1K	1K	1K	1K

Table 1: Statistics of the authentic (triplets contain human-post-edited translations) and synthetic APE datasets. For each language pair, the size of an existing and newly generated dataset is the same.

ing + Round-trip translation pipeline are closer to their corresponding *pe* sentences, as similar to Junczys-Dowmunt and Grundkiewicz (2016), we also maintain a weak connection between *pe* and *mt_new* sequence pairs. However, this may not always hold as the paraphrasing and round-trip translation processes are uncontrolled.

Thus, we use a slight variation of the corpus interleaving approach used by (Lee et al., 2022a) to mix the existing and the newly generated synthetic APE datasets. We compare the existing triplets with the newly generated ones and replace the poor-quality triplets with their counterparts. To make this decision, for each pair of triplets *((src, mt, pe), (src, mt_new, pe))*, we compare the edit distance between *mt-pe* with the *mt_new-pe* edit distance to decide whether to use only the newly generated or both the triplets, as shown in Equation 1.

$$mt = \begin{cases} mt, mt\_new & \text{if } |ter(mt, pe) - \mu| \leq 2\sigma \\ mt\_new & \text{otherwise,} \end{cases} \quad (1)$$

Where $ter(\cdot)$ denotes a function that computes TER (Snover et al., 2006) score, μ and σ are the mean and standard deviation of the TER scores between *mt-pe* pairs of the authentic APE corpus.

4 Experimental Details

This section describes the experimental setup for conducting the experiments across English-German (En-De), English-Russian (En-Ru), English-Marathi (En-Mr), English-Hindi (En-Hi), and English-Tamil (En-Ta) language pairs.

4.1 Datasets

For En-De, En-Ru, and En-Mr experiments, we use the authentic and synthetic datasets released through WMT21 (Akhbardeh et al., 2021), WMT19 (Chatterjee et al., 2019), and WMT22 (Bhattacharyya et al., 2022) APE shared tasks, respectively. Similarly, for En-Hi and En-Ta pairs, we use resources released through WMT24 QEAPE[2] shared subtask. Table 1 shows the sizes of synthetic and authentic datasets for each language pair in terms of the number of triplets. For the En-De pair, 4M triplets are randomly sampled from a subset of eSCAPE-NMT (Negri et al., 2018) corpus with TER between *mt* and *pe* pairs being less than 70. We refer to these synthetic corpora as *'existing'* synthetic corpora. We use the corresponding WMT development sets to evaluate the En-De, En-Mr, En-Hi, and En-Ta pairs. For the En-Ru pair, we use the WMT19 test set for evaluation.

We also use the BPCC (Gala et al., 2023) corpus for En-Mr, En-Hi, and En-Ta pairs during the first stage of our proposed synthetic data generation process. No additional parallel corpora are used for the En-De and En-Ru pairs.

4.2 Synthetic Data Generation

For generating En-Mr, En-Hi, and En-Ta synthetic APE data, we train NMT systems from scratch using the BPCC (Gala et al., 2023) corpus and using the set of *src-pe* pairs from existing synthetic data. For all three pairs, we use MultiIndicParaphraseGeneration (Kumar et al., 2022) model for paraphrasing with default configuration.

For En-De and En-Ru pairs, we do not use any additional parallel corpus apart from the *src-pe* from the existing synthetic APE datasets. Instead of training a model from scratch, we fine-tune the NLLB-1.3B (Costa-jussà et al., 2022) model for translation into both directions. For paraphrasing the German reference sentences, we use *milyiyo/paraphraser-german-mt5-small*[3] with default configuration. Similarly, for paraphrasing Russian sentences, we use the *cointegrated/rut5-base-paraphraser*[4] with the default configuration. We use *eugenesiow/bart-paraphrase*[5] for English with the default configuration. The training approach is discussed in Appendix 4.5.

4.3 Experiments

This section describes experiments performed using existing and newly generated synthetic APE datasets. Each experiment involves training an APE model using the same training approach. Descriptions of each experiment are as follows.

Do Nothing This baseline considers original translations as APE outputs.

[2] WMT24 QEAPE Shared Subtask
[3] German Paraphraser
[4] Russian Paraphraser
[5] English Paraphraser

Existing Synthetic We do not use the newly generated synthetic APE corpus for this experiment. Rather, the APE models are trained using the existing synthetic data, which is generated using the approach proposed by Negri et al. (2018). We consider this experiment as our **primary baseline**.

Proposed Synthetic In this experiment, the newly generated synthetic APE data is used in place of the existing synthetic APE data.

Half Existing + Half Proposed This experiment uses randomly sampled half of the existing synthetic APE data and the other half of the newly generated synthetic data triplets. Thus, *src* and *pe* are not repeated.

Existing OR Proposed Each pair of triplets from the existing and newly generated synthetic APE data is taken, and one with the lower TER score between translation and post-edit is added to the final synthetic dataset.

Existing + Proposed Synthetic Both synthetic corpora are used in training the APE models.

Corpus Interleaving In this experiment, the existing and newly generated synthetic corpora are mixed as per Equation 1.

4.4 Training Details

To ensure consistency across all experiments, we maintain a uniform set of configurations. Our APE models undergo training with a batch size of 32, and we use early stopping with a patience of 5. A maximum of 5000 epochs are allowed. The Adam optimizer is used with a learning rate of 5 x 10^{-5}, and β_1 and β_2 are set to 0.9 and 0.997, respectively. Additionally, we use 15,000 warmup steps. We use the beam search with a beam size of 5 for decoding. All the experiments are carried out using NVIDIA A100 GPUS. The APE model contains around 40M parameters, and training the model using CTS requires about 48 hours. As the time required for each experiment is and we have performed multiple experiments over each of the five language pairs, we report single-run results. For pre-processing the English, German, and Russian data, we use the NLTK library[6], while the Indic-NLP library[7] is used for processing Marathi, Hindi, and Tamil text. We used Pytorch[8] for Model training and inference. For computing the TER scores, we use the official WMT APE evaluation script[9], and for computing the BLEU scores, we use the SacreBLEU[10] library.

4.5 Training Approach

We follow the Curriculum Training Strategy (CTS) similar to Deoghare and Bhattacharyya (2022) for training our APE systems. It involves gradually adapting a model to more and more complex tasks. The steps of the CTS are described below.

In the first step, we train a model for the APE task using the synthetic APE data in the two phases. In the first phase, we train the model over a subset of the synthetic corpus containing triplets with poorer TER than the *Do Nothing* baseline. In the second phase, we train the model over the other subset of the synthetic corpus, containing triplets with equal or better TER than the *Do Nothing* baseline. Finally, we fine-tune the APE model using in-domain authentic APE data. The training details are described in subsection 4.4.

5 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. Additionally, we report BLEU (Papineni et al., 2002) scores for the same experiments in Appendix C. We perform a statistical significance test considering the primary metric (TER) using William's significance test (Graham, 2015).

While we primarily performed all experiments as single-run experiments due to the large number of experiments and limited access to the compute resource, to check whether the experiments were very sensitive to the parameter initialization, we report mean TER scores over two runs of each experiment for English-Marathi and English-Hindi pairs in Appendix D (Refer Table 6).

Case Analysis Figure 2 shows two En-Hi examples from the existing and the newly generated synthetic APE datasets. In both the examples, *Existing Translation* and *New Translation* refer to translations in the existing and the newly generated synthetic datasets, respectively. The Source and Reference sentences are from their respective triplets.

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

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

[8] https://pytorch.org/

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

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

Example 1		
Source ***(src)***	Check out silent heart attack symptoms and causes	
Existing Translation ***(mt)***	दिल का दौरा पड़ने के लक्षण और कारण	Dil *(Heart)* ka *(of)* dauraa *(attack)* padane *(having)* ke *(of)* lakshan *(symptoms)* aur *(and)* kaaran *(cause)*
Reference ***(pe)***	साइलेंट हार्ट अटैक के लक्षण और कारण देखें	Silent *(Silent)* hart *(heart)* attack *(attack)* ke *(of)* lakshan *(symptoms)* aur *(and)* kaaran *(causes)* dekhen (check)
New Translation ***(mt_new)***	जानें साइलेंट हार्ट अटैक के लक्षण और कारण	*Jaanen (check) sayilent (silent) hart (heart) attack (attack) ke (of) lakshan (symptoms) aur (and) kaaran (causes)*
Example 2		
Source ***(src)***	Depression is a highly prevalent and disabling disease.	
Existing Translation ***(mt)***	अवसाद एक अत्यधिक प्रचलित और अक्षम बीमारी है।	Awasaad *(Depression)* ek *(one)* atyadhik *(highly)* prachalit *(prevalent)* aur *(and)* aksham *(disabling)* bimaari *(disease)* hai *(is)* |
Reference ***(pe)***	अवसाद एक अत्यधिक प्रचलित और अक्षम कर देने वाली बीमारी है।	Awasaad (Depression) ek *(one)* atyadhik *(highly)* prachalit *(prevalent)* aur *(and)* aksham *(disabling)* kar *(does)* dene *(give)* waali *(which)* bimaari *(disease)* hai *(is)* |
New Translation ***(mt_new)***	अवसाद एक खतरनाक और खतरनाक बीमारी है।	Awasaad *(Dipression)* ek *(one)* khataranaak *(dangerous)* aur *(and)* khataranaak *(dangerous)* bimaari *(disease)* hai *(is)* |

Figure 2: Example APE triplets from the newly generated English-Hindi synthetic data along with their corresponding translations from the existing synthetic APE data.

Experiment	En-De	En-Ru	En-Mr	En-Hi	En-Ta
Do Nothing	19.06	16.16	22.93	44.36	28.34
Existing Synthetic	18.71	15.97	19.01	22.41	21.00
w/o Paraphrasing + w/ RTT	19.00	16.04	19.69	24.91	23.02
w/ Paraphrasing + w/o RTT	18.68	16.11	18.93	23.00	21.84
w/ Paraphrasing + w/ RTT	18.49	**15.78**	**18.66**	**20.83**	**19.98**
BT + w/ Paraphrasing + FT	**18.44**	15.93	18.80	20.91	20.78

Table 2: TER scores of APE models, on their respective evaluation sets, trained on the newly generated synthetic data with and without paraphrasing and roundtrip-translation (RTT). The *w/o Paraphrasing + w/ RTT* experiment follows the same synthetic APE data generation approach as Freitag et al. (2019).

In the first example, the *mt* contains translations of all phrases except 'silent' and 'check out.' The *pe* sentence contains the transliteration of the phrase 'silent heart attack,' along with the correct translation of all other phrases in the source. Therefore, using *mt* as a translation would lead to the APE model learning to modify even the correctly translated words. On the other hand, the *mt_new* contains the transliteration of the 'silent heart attack' phrase. Thus reducing the edit distance between the translation and post-edit. Also, in this case, the *mt_new* is a reordered version of the *pe*, which some may find relatively less fluent.

The second example shows how using the proposed method for synthetic data generation may lead to a very poor quality translation generation. The *mt* in this example exhibits an adequacy issue as the sentence means 'the disease is disabling.' Yet, correct translations of all words in the source sentence are present. Merely an insertion of '*kar dene waali*' phrase can fix this. However, the *mt_new* is even worse where both 'prevalent' and 'disabling' phrases are translated separately as '*khatarnaak*,' which means 'dangerous.' Also, the word 'highly' is not translated explicitly. Correcting this translation would require multiple edit operations. Thus, we conjecture that such triplets are the reason that the corpus interleaving helps improve the APE performance.

Impact of Paraphrasing Table 2 shows the results of the APE models trained using the newly generated synthetic APE data, with and without paraphrasing and round-trip translations.

Outcomes of this comparison suggest that if the *pe* sentences are not paraphrased and directly fed to MT systems for a round-trip translation, the generated *mt_new* sentences do not reflect the translation errors as in the authentic APE data. Merely performing round-trip translation is likely to generate translations similar to their references as the same parallel corpus is used for training both the NMT

models.

Significant performance improvement when paraphrasing is used suggests that the modifications done by a paraphraser bring more feature diversity. Further, the comparison between *w/o Paraphrasing + w/ RTT* and *w/ Paraphrasing + w/ RTT* experiments show the importance of paraphrasing in synthetic data generation. The paraphraser provides the target-to-source NMT system with a sentence with enough variation that the final erroneous translation serves as a good candidate for the APE triplet.

Also, the comparable results of *w/ Paraphrasing + w/o RTT* with *Existing Synthetic* show performing paraphrasing alone does lead to performance improvements. Further exploration of the paraphrasing impact is discussed in Appendix E.

The *w/ Paraphrasing + w/o RTT* experiment performs paraphrasing of the back-translation of the original reference, which is in English, and then the forward translation is generated to get the alternative of the original reference sentence. The small performance difference between this experiment and the *w/ Paraphrasing + w RTT* experiment shows performing the paraphrasing on the possibly erroneous source sentence is less effective than paraphrasing the original reference sentence.

Quantitative Analysis Table 3 compiles TER scores achieved in various experiments on WMT21 and WMT22 development and WMT19 test sets by En-De, En-Mr, and En-Ru APE systems, respectively. TER scores on in-house created test sets are reported for En-Hi and En-Ta APE experiments. Along with the in-house-created training and development APE data, these test sets will also be released through the upcoming WMT shared task.

The first two rows of Table 3 show results for the baselines. *Do Nothing* baseline treats original translations as APE outputs. *Existing Synthetic* baseline uses the existing synthetic data in training the APE model. We regard *Existing Synthetic* as a **primary baseline** for this work.

Proposed Synthetic, *Half Existing + Half Proposed*, and *Existing OR Proposed* experiments utilize an equal amount of synthetic and authentic APE data. Gains observed in the *Proposed Synthetic* experiment over the *Existing Synthetic* baseline regarding TER scores across all language pairs denote the better quality of the newly generated synthetic data.

The slightly poorer performance observed in the *Half Existing + Half Proposed* experiment again suggests that the newly generated synthetic APE data is more beneficial than the existing one. However, to concretize this observation, running similar experiments by drawing out multiple random half-sized subsets from the existing synthetic data and then selecting the corresponding other half subsets from the newly generated data is required.

We carried out the *Existing OR Proposed* experiment to investigate whether naively selecting either of the triplets with the same pair of *src* and *pe* sequences from the existing and the newly generated data helps APE systems improve their performance. For En-De, En-Ru, En-Hi, En-Mr, and En-Ta pairs, about 73%, 59%, 80%, 84%, and 77% triplets from the proposed method are chosen, respectively. Though insignificant, based on the slight improvements over the corresponding primary baselines, we can say that the proposed synthetic data generation technique is not highly prone to generating poor-quality *mt_new* sentences. Otherwise, the results would have been closer to that of the *Existing Synthetic* experiment. A possible reason for this could be the use of the same *src* and *pe* pairs from the existing corpus to train the NMT models. This comparison also highlights that the triplets with closer translation and *pe* sequences are important for improving APE models.

As the generation of *mt_new* sequences is uncontrolled, there could also be triplets with poor-quality translations in the newly generated synthetic data. Similarly, not all triplets in the existing synthetic data could be deemed unfit for training APE models. Furthermore, as seen by Yu et al. (2023) and Deoghare and Bhattacharyya (2022), exposing additional translation for the same *src* sentence leads to performance enhancements. Thus, we conduct *Existing + Proposed Synthetic* and *Corpus Interleaving* experiments.

Simply augmenting the newly generated data with the existing one (*Existing + Proposed Synthetic*) results in improvements over the primary baselines for all language pairs except En-Ru. This underlines that having multiple triplets with the same *src* and *pe* but different translations helps improve APE performance. Further, adding a filter (*Corpus Interleaving*) that filters out triplets from the existing synthetic APE corpus, which have a high edit distance between translation and *pe* pairs, results in additional small improvements in performance for some of the pairs. Appendix F discusses a variation of the *Corppus Interleaving* experiment,

Experimental Setting	En-De	En-Ru	En-Mr	En-Hi	En-Ta
Do Nothing	19.06	16.16	22.93	44.36	28.34
Existing Synthetic	18.71	15.97	19.01	22.41	21.00
Proposed Synthetic	18.49*	15.78	18.66	20.83	19.98
Half Existing + Half Proposed	18.69*	15.86*	18.73	21.05	20.60
Existing OR Synthetic	18.40	15.74	18.57	20.77	19.96
Existing + Proposed Synthetic	**18.33**	16.05	18.29	20.30	18.65
Corpus Interleaving	18.35	15.81*	**18.00**	**19.99**	**18.40**
Junczys-Dowmunt and Grundkiewicz (2016)	19.31	16.10	21.38	23.02	21.93
Lee et al. (2020)	19.28	16.02	20.95	22.29*	20.90*
Tuan et al. (2021)	19.10	16.09	19.04	22.20*	20.77*
Lee et al. (2022a)	18.35	15.82	18.96	21.34	19.41
Lee et al. (2022b)	18.42	**15.71**	18.88	21.07	19.23

Table 3: TER scores of the APE systems on the respective evaluation sets. All models are trained using the Curriculum Training Strategy over synthetic and authentic APE data. The rows after the *Corpus Interleaving* experiment report results of the experiments ran using the existing synthetic APE data generation approaches. * denotes that the improvement over the primary baseline (Existing Synthetic) is insignificant (p being 0.05). Refer to Appendix C for BLEU scores.

which also filters triplets from the newly generated data.

Smaller gains for the language pairs with tough *Do Nothing* baselines highlight the difficulty in developing APE systems that can precisely identify and correct translation errors.

Comparison with Existing Synthetic Data Generation Methods We also compare our proposed synthetic data generation method with other synthetic APE data generation methods discussed in the Related Work section. Due to inconsistencies in data, modeling, training, and decoding techniques used across these works, we have only used their synthetic APE data generation techniques and kept other components consistent with the experiments in this work. We have reported the result after using *Corpus Interleaving* with each synthetic data generation technique to report the best results. The last five rows of Table 3 show except for En-Ru, our proposed synthetic APE data generation technique outperforms earlier proposed methods. For En-De and En-Ru pairs, we observe comparable performances of our proposed technique and of Lee et al. (2022b).

6 Conclusion and Future Work

In this work, we presented a reference-based synthetic APE data generation technique that uses paraphrasing and round-trip translation to obtain a corrupted translation from a reference sentence in a parallel corpus. Except for the target language paraphraser, our technique does not require any other linguistic resources of the source or target language. Experimental results across En-De, En-Ru, En-Mr, En-Hi, and En-Ta pairs, which are from different language families, show that using the synthetic data generated through the proposed method in training APE models results in better performance than the corresponding APE models that use the existing synthetic data during their training. Also, we observe the best performance when the newly generated synthetic data is augmented with existing synthetic data. Our En-Mr APE model achieves the state-of-the-art performance on the WMT22 development set. We will release the newly generated synthetic APE data under the CC-BY-SA 4.0 license publicly for further research. In the future, we wish to explore in detail how a choice of paraphraser used during erroneous translation generation impacts the triplet quality. It will help in working towards a controlled generation of erroneous translations.

7 Limitations

Similar to the existing approach, our approach too relies on the use of a parallel corpus. Therefore, the amount of synthetic APE data that can be generated through the proposed method is limited by the size of the available parallel corpus. Also, though our approach does not rely on the linguistic features of the source and target languages, it does require a paraphraser for the target language. In this work,

we have not explored how the quality of the paraphraser affects the corrupted translations. Since the generation of corrupted sentences through paraphrasing and round-trip translation is uncontrolled, our approach can produce poor-quality translations. It leads to the requirement of using some technique to decide whether to use the generated erroneous translation or not. Furthermore, even though the current work already covers five different language pairs, such uncontrolled generation limits us from estimating the generalizability of this technique to other languages.

8 Ethics Statement

Our APE models are trained 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. Furthermore, these are standard benchmarks released through 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 generated as a part of this work will be released under the CC-BY-SA 4.0 license publicly 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

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.

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.

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*.

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.

Elozino Egonmwan and Yllias Chali. 2019. Transformer and seq2seq model for paraphrase generation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 249–255, Hong Kong. Association for Computational Linguistics.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. APE at scale and its implications on MT evaluation biases. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

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.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2016. Log-linear combinations of monolingual and bilingual neural machine translation models for automatic post-editing. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 751–758, Berlin, Germany. Association for Computational Linguistics.

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.

Aman Kumar, Himani Shrotriya, Prachi Sahu, Amogh Mishra, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Mitesh M. Khapra, and Pratyush Kumar. 2022. IndicNLG benchmark: Multilingual datasets for diverse NLG tasks in Indic languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5363–5394, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

WonKee Lee, Seong-Hwan Heo, Baikjin Jung, and Jong-Hyeok Lee. 2022a. Towards semi-supervised learning of automatic post-editing: Data-synthesis by infilling mask with erroneous tokens. *ArXiv*, abs/2204.03896.

WonKee Lee, Baikjin Jung, Jaehun Shin, and Jong-Hyeok Lee. 2021. Adaptation of back-translation to automatic post-editing for synthetic data generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3685–3691, Online. Association for Computational Linguistics.

Wonkee Lee, Baikjin Jung, Jaehun Shin, and Jong-Hyeok Lee. 2022b. Reshape: Reverse-edited synthetic hypotheses for automatic post-editing. *IEEE Access*, 10:28274–28282.

WonKee Lee, Jaehun Shin, Baikjin Jung, Jihyung Lee, and Jong-Hyeok Lee. 2020. Noising scheme for data augmentation in automatic post-editing. In *Proceedings of the Fifth Conference on Machine Translation*, pages 783–788, Online. Association for Computational Linguistics.

Jessy Lin, Geza Kovacs, Aditya Shastry, Joern Wuebker, and John DeNero. 2022. Automatic correction of human translations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 494–507, Seattle, United States. Association for Computational Linguistics.

Hyeonseok Moon, Chanjun Park, Seolhwa Lee, Jaehyung Seo, Jungseob Lee, Sugyeong Eo, and Heuiseok Lim. 2022a. Empirical analysis of noising scheme based synthetic data generation for automatic post-editing. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 883–891, Marseille, France. European Language Resources Association.

Hyeonseok Moon, Chanjun Park, Jaehyung Seo, Sugyeong Eo, and Heuiseok Lim. 2022b. An automatic post editing with efficient and simple data generation method. *IEEE Access*, 10:21032–21040.

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).

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.

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.

Yi-Lin Tuan, Ahmed El-Kishky, Adithya Renduchintala, Vishrav Chaudhary, Francisco Guzmán, and Lucia Specia. 2021. Quality estimation without human-labeled data. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 619–625, Online. 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 *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jiayi Wang, Ke Wang, Kai Fan, Yuqi Zhang, Jun Lu, Xin Ge, Yangbin Shi, and Yu Zhao. 2020. Alibaba's submission for the WMT 2020 APE shared task: Improving automatic post-editing with pre-trained conditional cross-lingual BERT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 789–796, Online. Association for Computational Linguistics.

Jiawei Yu, Min Zhang, Zhao Yanqing, Xiaofeng Zhao, Yuang Li, Su Chang, Yinglu Li, Ma Miaomiao, 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*, pages 926–930, Singapore. Association for Computational Linguistics.

Richard Zens, Franz Josef Och, and Hermann Ney. 2002. Phrase-based statistical machine translation. In *KI 2002: Advances in Artificial Intelligence: 25th Annual German Conference on AI, KI 2002 Aachen, Germany, September 16–20, 2002 Proceedings 25*, pages 18–32. Springer.

A Post-editing Data Sample

English-Marathi Example	
Source	The children enjoyed playing in the park during the evening.
Reference from Parallel Corpus	मुलांनी संध्याकाळी बागेत खेळण्याचा भरपूर आनंद लुटला.
MT	मुलांनी संध्याकाळी उद्यानामध्ये खेळायला आनंद घेतला होता.
MT + Human Post-Edited	मुलांनी संध्याकाळी उद्यानात खेळण्याचा आनंद लुटला.
MT + APE	मुलांनी संध्याकाळी उद्यानामध्ये खेळण्याचा आनंद घेतला.

Figure 3: An example from the English-Marathi pair

Figure 3 contains a representative English-Marathi example that shows that independently obtained reference and translation are more distant than the translation and its post-edited version. Similarly, we see independently obtained references, and the post-edited machine-translated sentences are distant in terms of TER score, too. It shows why it is not a good strategy to treat a reference sentence from a parallel corpus as a reference post-edit for independent MT-generated translation.

B Background: Automatic Post-Editing

This section describes the learning objective and architecture of an APE model.

Learning Objective Considering the aim of APE to identify and correct erroneous *mt* and generate *pe* by ensuring that the meaning of *src* is preserved, both *src* and *mt* sentences are crucial. *src* is considered as an auxiliary sequence that gives necessary contextual information and helps spot translation errors. While *mt* serves as the primary sequence which needs to be corrected. Considering *src*, *mt* and *pe* as $x = \{x_i\}_{i=1}^{T_x}$, $y = \{y_i\}_{i=1}^{T_y}$, and $z = \{z_i\}_{i=1}^{T_z}$ with lengths T_x, T_y, and T_z, respectively, the APE model learns to generate *pe* with the conditional probability as shown in Equation 2.

$$p(z) = \prod_{k=1}^{T_z} p(z_k | x, y, z_{<k}; \Theta) \quad (2)$$

Where θ denotes the model parameters.

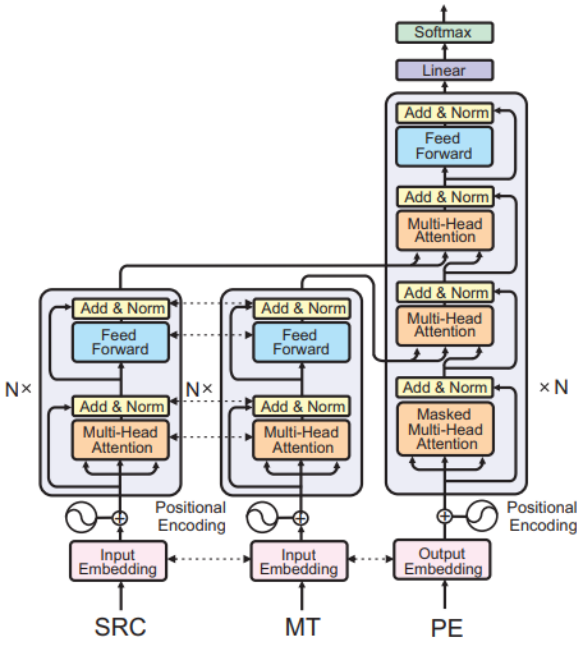

Figure 4: Dual-Encoder Single-Decoder APE Architecture. Dashed arrows represent tied parameters and common embedding matrices for encoders and for the decoder (Deoghare and Bhattacharyya, 2022).

Equation 3 shows the cross-entropy loss function used to train the APE model.

$$L_{APE} = -\sum_{w=1}^{|S|}\sum_{e=1}^{|V|} y_{w,e} \log\left(\hat{y}_{w,e}\right) \quad (3)$$

Where $|S|$ and $|V|$ represent the number of tokens in sequence and the number of vocabulary tokens, respectively. The APE output is represented by the $\hat{y}_{w,e}$, and $y_{w,e}$ denotes ground truth and prediction, respectively.

Architecture We construct the APE system using a transformer-based encoder-decoder framework. For English-Indian language APE systems, we employ two distinct encoders to process a source sentence and its corresponding translation, given the absence of script or vocabulary overlap between these languages. The outputs from both encoders are then passed to two successive cross-attention layers within the decoder.

The architecture depicted in Figure 4 illustrates the architecture of the English-Indian language APE model. A single-encoder single-decoder architecture is utilized for the English-German and English-Russian APE due to the linguistic and lexical similarities between these languages. A single encoder is responsible for encoding the concatenation of the source and translated text, which

Experiment	En-De	En-Ru	En-Mr	En-Hi	En-Ta
Do Nothing	68.79	76.20	64.51	39.53	67.55
Existing Synthetic	69.17	76.35	68.76	63.04	75.58
Proposed Synthetic	69.26	76.53	69.19	64.79	76.65
Half Existing + Half Proposed	69.20	76.47	69.11	64.55	76.00
Existing OR Synthetic	69.35	76.56	69.28	64.86	76.66
Existing + Proposed Synthetic	69.71	76.29	69.74	65.38	77.80
Corpus Interleaving	**69.74**	76.52	**69.98**	**65.67**	**78.02**
Junczys-Dowmunt and Grundkiewicz (2016)	68.47	76.25	66.22	65.23	74.68
Lee et al. (2020)	68.48	76.38	75.90	77.11	75.65
Tuan et al. (2021)	68.69	76.25	68.81	66.96	75.47
Lee et al. (2022a)	69.66	76.50	18.96	21.34	76.57
Lee et al. (2022b)	69.60	**76.58**	18.88	21.07	76.73

Table 4: BLEU scores of the APE systems on the respective evaluation sets. All models are trained using the Curriculum Training Strategy over synthetic and authentic APE data.

Experiment	En-De	En-Ru	En-Mr	En-Hi	En-Ta
Do Nothing	19.06	16.16	22.93	44.36	28.34
Existing Synthetic	18.71	15.97	19.01	22.41	21.00
t = 1	18.49	15.78	18.66	20.83	19.98
t = 0.75	**18.36**	15.76	18.58	20.79	19.85
t = 0.5	18.38	**15.74**	**18.49**	**20.71**	**19.82**
t = 0.25	18.51	15.76	18.54	20.75	19.94
t = 0	18.50	15.85	18.60	21.00	20.32

Table 5: TER scores of APE models on their respective evaluation sets when trained on the newly generated synthetic data with different temperature (**t**) values used during the paraphrase generation.

is done by inserting a '<*SEP*>' token between them. Subsequently, the encoder output is passed to a single cross-attention layer within the decoder. For all pairs, the encoders are initialized with IndicBERT (Kakwani et al., 2020) weights.

C Quantitative Evaluation: BLEU Scores

Table 4 shows the BLEU scores for the same experiments for which the TER scores are reported in Table 3.

D Impact of Randomness in Parameter Initialization

Due to the limited access to compute resources, experiments performed in this work are single-run experiments. Though we carefully choose the hyperparameters so that the models converge well, different initial parameter initializations may lead to different results.

In order to investigate the impact of randomness, we perform two runs of each experiment for English-Marathi and English-Hindi language pairs.

Experimental Setting	En-Mr	En-Hi
Do Nothing	22.93	44.36
Existing Synthetic	19.04	22.42
Proposed Synthetic	18.65	20.80
Half Existing + Half Proposed	18.73	21.07
Existing OR Synthetic	18.50	20.77
Existing + Proposed Synthetic	18.35	20.24
Corpus Interleaving	**17.94**	**20.03**

Table 6: Mean TER scores computed over two runs of each experiment of English-Marathi and English-Hindi language pairs whose results are reported in Table 3. * denotes that the improvement over the primary baseline (Existing Synthetic) is insignificant (p being 0.05).

Each run of an experiment uses the same data and the same hyperparameters. Table 6 reports the mean TER score (primary metric) for all experiments.

The results reveal the same pattern as visible in Table 3.

Experimental Setting	En-De	En-Ru	En-Mr	En-Hi	En-Ta
Do Nothing	19.06	16.16	22.93	44.36	28.34
Proposed Synthetic	18.49*	15.78	18.66	20.83	19.98
Existing + Proposed Synthetic	**18.33**	16.05	18.29	20.30	18.65
Corpus Interleaving (Data Augmentation)	18.35	15.81*	**18.00**	**19.99**	**18.40**
Corpus Interleaving (Data Selection)	18.49	16.03	18.12	20.16	18.47

Table 7: TER scores of the APE systems on the respective evaluation sets.

E Impact of Quality of Paraphrases

Table 5 shows the TER scores on the respective evaluation sets when different temperature values are used. From the results, we conjecture that the intensity of the impact of the quality of the paraphrased generation will be reduced as we ultimately fine-tune the model using the authentic APE data.

F Impact of Quality of Paraphrases

Table 7 shows the comparison between the data-augmentation-based *Corpus Interleaving* experiment, referred here as *Corpus Interleaving (Data Augmentation)*, discussed in Section 3. Unlike the *Corpus Interleaving (Data Augmentation)* experiment, the *Corpus Interleaving (Data Selection)* also filters triplets from the newly generated data.