Advancing Semi-Supervised Learning for Automatic Post-Editing: Data-Synthesis by Mask-Infilling with Erroneous Terms

Wonkee Lee[12*]**, Seong-Hwan Heo**[2]**, Jong-Hyeok Lee**[2]

[1]LG AI Research, [2]POSTECH

wklee@lgresearch.ai, {hursung1, jhlee}@postech.ac.kr

Abstract

Semi-supervised learning that leverages synthetic data for training has been widely adopted for developing automatic post-editing (APE) models due to the lack of training data. With this aim, we focus on data-synthesis methods to create high-quality synthetic data. Given that APE takes as input a machine-translation result that might include errors, we present a data-synthesis method by which the resulting synthetic data mimic the translation errors found in actual data. We introduce a noising-based data-synthesis method by adapting the masked language model approach, generating a noisy text from a clean text by infilling masked tokens with erroneous tokens. Moreover, we propose selective corpus interleaving that combines two separate synthetic datasets by taking only the advantageous samples to enhance the quality of the synthetic data further. Experimental results show that using the synthetic data created by our approach results in significantly better APE performance than other synthetic data created by existing methods.

Keywords: Automatic post-editing, synthetic-corpus generation, masked language modeling

1. Introduction

Automatic post-editing (APE) (Chatterjee et al., 2015, 2018) is a study of correcting errors in machine translation (MT) outputs to provide high-quality (publishable) translations. APE can be conceptually framed as a multi-source sequence-to-sequence problem, as depicted in Figure 1, which concurrently takes both a source text (src) and its machine-translated text (mt) to generate a post-edited text (pe). This process can be succinctly represented as a mapping of $(src, mt) \rightarrow pe$, and is typically addressed through supervised learning methods.

Accordingly, APE models necessitate the use of triplet data (src, mt, pe), commonly referred to as an **APE triplet**, for supervised learning. This data is established on the underlying assumption (Bojar et al., 2015) that pe serves as the **minimum correction** of mt, and both src and pe are assumed to be **free from error** and **semantically equivalent**. Despite the potential of APE models, the scarcity of high-quality human-made (gold) APE data poses a significant challenge in the development of robust APE models. To overcome this problem, semi-supervised learning, which leverages synthetic data in addition to gold data for training, has been instrumental in building APE models.

The central focus of this approach lies in the data-synthesis method, which aims to generate synthetic data that meets both quality and quantity requirements. Specifically, the resulting synthetic data should possess sufficient volume and accurately reflect the inherent characteristics found in the gold data, such as statistical properties. For this pur-
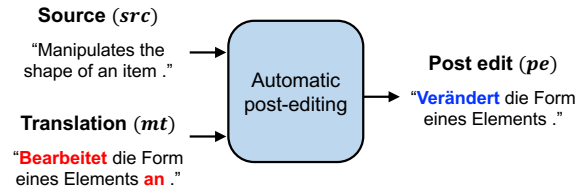

Figure 1: Overview of the APE process for an English-German translation. Bold highlights indicate erroneous words in mt and post-edited words in pe.

pose, numerous studies (Junczys-Dowmunt and Grundkiewicz, 2016; Negri et al., 2018; Lee et al., 2020, 2021) have proposed methods to generate synthetic APE data, including the methods (Negri et al., 2018; Lee et al., 2020, 2021) that produce synthetic APE data using a bilingual parallel corpus, which have received significant attention (detailed in Section 2). Although these have been proven beneficial to APE training, their synthetic data have drawbacks in that (1) the included translation errors excessively outnumber those in gold data (Negri et al., 2018; Lee et al., 2021) or (2) the errors are not likely to appear in gold data (Lee et al., 2020).

In this study, we introduce a data-synthesis approach through the application of a noising mechanism to the masked language model (MLM) framework (Devlin et al., 2019; Conneau and Lample, 2019). This technique, we have termed **MLM noising**, aims to address and mitigate the shortcomings prevalent in the existing data synthesis methods. The objective of employing this method is to ensure that the resulting synthetic data exhibit translation errors that closely replicate those found in gold data, both in quality and quantity. Specifically, we

---

*Work completed while the author was a student at POSTECH.

first manipulate a clean target text (i.e., a reference text in a parallel corpus) by replacing a specific portion with `[MASK]` tokens according to the error statistics from gold data and then let our MLM substitute each `[MASK]` with an **erroneous token** that is likely to appear in gold data. We note that filling masks with erroneous tokens deviates from the standard approach of MLM, which aims to predict the correct tokens.

Furthermore, we contend that our MLM noising method and the existing method (Negri et al., 2018) may offer complementary benefits due to their orthogonal advantages. To leverage the strengths of both approaches, we introduce a corpus ensemble method **selective corpus interleaving**. This method selectively incorporates advantageous samples from two synthetic datasets, choosing those that most closely resemble the gold data. We experimentally demonstrated that our synthetic data significantly improves the model performance compared to other synthetic data created by existing methods.

In summary, our key contributions to this work are as follows:

- **MLM noising:** We introduce MLM noising, a method that adapts the masked language model (MLM) for creating synthetic APE data. This approach strategically inserts `[MASK]` tokens into clean target texts and replaces them with unnatural tokens that are likely to appear as translation errors in gold data.

- **Selective Corpus Interleaving:** We present a corpus ensemble strategy designed to combine two distinct synthetic datasets. This method makes an integrated dataset by carefully selecting the most advantageous samples from two different datasets that better resemble the characteristics of gold data.

2. Background

2.1. Problem Statement

APE has been mainly framed as a dual-source sequence-to-sequence learning problem $(src, mt) \rightarrow pe$, taking two inputs: src, which is responsible for providing contextual information to identify translation errors, and mt, which is the object of correction. Formally, let $D = \{(\mathbf{x}, \widetilde{\mathbf{y}}, \mathbf{y})_i\}_{i=1}^{n}$ denotes a set of n APE triplets (whether they are gold or synthetic triplets), where $\mathbf{x} = (x_i \ldots x_{T_x})$, $\widetilde{\mathbf{y}} = (\widetilde{y}_1 \ldots \widetilde{y}_{T_{\widetilde{y}}})$, and $\mathbf{y} = (y_1 \ldots y_{T_y})$ indicate src, mt, and pe, respectively[1]. An APE model learns to the estimate parameters θ that maximize the

---

[1]These notations will be applied across all mathematical formulations presented in this paper.
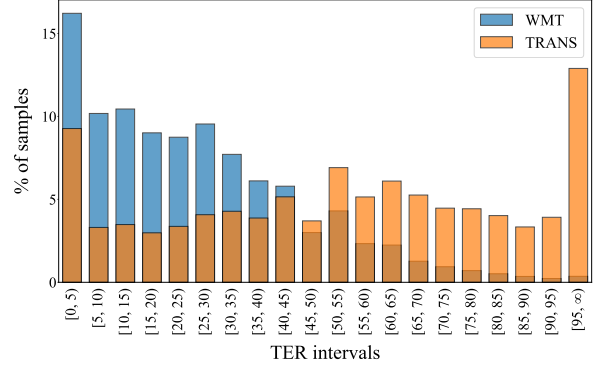

Figure 2: Categorical distributions of gold (WMT) and synthetic (TRANS) APE data, representing the proportion of samples belonging to a particular interval of translation error rate (TER) [%], a similarity measure based on the edit distance between mt and its target (i.e., pe or ref).

conditional probability on pe as:

$$P(\mathbf{y}|\mathbf{x}, \widetilde{\mathbf{y}}) \approx \arg\max_{\theta} \sum_{i=1}^{T_y} \log P_{\theta}(y_i \mid y_{<i}, \mathbf{x}, \widetilde{\mathbf{y}}). \quad (1)$$

2.2. Existing Data-Synthesis Methods

In this section, we outline existing methods that employ a **bilingual parallel corpus** to create synthetic APE triplets. In section 2.2.1, we provide an overview of an early method that continues to be widely used. Subsequently, in sections 2.2.2 and 2.2.3, we discuss subsequent studies that have expanded this foundational approach.

2.2.1. Translation Approach (TRANS)

Translation approach (TRANS), introduced in an early study by Negri et al. (2018), motivates the generation of synthetic APE data using a parallel corpus. This method remains the predominant method used in many studies. Given that the parallel corpus consists of a source (src) and the corresponding reference translation (ref) pairs, namely, (src, ref), this method configures synthetic APE triplets in the form of (src, mt, ref), in which src and ref are taken from the parallel corpus and mt is the machine translation result of src.

Despite its simplicity, this method has an advantage in mirroring the attributes of gold data, potentially leading to consistency in quality with gold data. Specifically, src and ref in the parallel corpus are supposed to be error-free and have identical semantics, replicating the relationship observed between src and pe in gold APE data. Additionally, mt is generated following the same process as applied to gold data.

However, a critical limitation of this approach is the lack of assurance that the ref (serving as a

proxy for pe) represents the minimal correction required for mt. This is because ref is generated independently of mt—that is, ref is not created by directly correcting mt—thereby deviating from the fundamental principle that gold APE data should follow. Consequently, the synthetic data may exhibit a significantly higher number of errors compared to gold data, potentially prompting the APE model to apply overly aggressive corrections to mt. This discrepancy, as demonstrated by the differing edit-distance distributions between $mt\text{-}ref$ and $mt\text{-}pe$ (Figure 2), indicates considerable potential for enhancing the quality of the method. By addressing this discrepancy and more closely aligning with the characteristics of gold data, we anticipate the production of higher-quality synthetic data.

2.2.2. Back-Translation Approach (BT)

This approach (Lee et al., 2021) modifies the TRANS triplet (src, mt, ref) by replacing the machine-translated text (mt) with a synthesized counterpart ($\widetilde{mt}$), thus yielding the new configuration: $(src, \widetilde{mt}, ref)$. This is accomplished through an adaptation of the back-translation (BT) technique, which traditionally generates synthetic source texts from target texts, by incorporating the APE process. This method has proven effective in reducing the edit distance to ref when $\widetilde{mt}$ is used instead of mt, thereby mitigating the problem encountered by the TRANS method. This method introduced two approaches for producing $\widetilde{mt}$:

- Forward generation (BT-FG): This approach involves applying a (src, mt) pair to the APE process to create a partially corrected version ($\widetilde{mt}$), aiming for reduced distance to ref.

- Backward generation (BT-BG): This approach applies a (src, ref) pair to the reversed APE process, $(src, pe) \rightarrow mt$, to produce $\widetilde{mt}$ that likely contains translation errors, presenting it as a partially degraded reference.

However, the degree of reduction in the edit distance achieved by using their $\widetilde{mt}$ is insignificant, so the discrepancy remains, and moreover, the resulting synthetic data themselves improve the APE performance less than TRANS (Section 4.3).

2.2.3. Random Noising Approach (RAND)

As in the BT approach, this method (Lee et al., 2020) substitutes mt of TRANS triplet with its synthetic machine translation $\widetilde{mt}$. This method introduces random noises (RAND) into ref to create $\widetilde{mt}$, aligning the number of introduced noises with the error distribution observed in gold data. Specifically, for each token in ref, the method probabilistically applies one of four editing operations—*keep*, *insert*,

delete, or *substitute*—reflecting translation errors. This selection is based on the categorical probability distribution of these operations derived from gold data, yielding a collection of synthetic APE triplets $(src, \widetilde{mt}, ref)$.

The primary advantage of this approach is its capacity to produce synthetic data whose error frequency mirrors that of gold data. However, a significant challenge arises from the qualitative disagreement between the synthetic and gold data, caused by the stochastic nature of the noising process. This randomness is evident in the procedures of word selection for insertion and substitution, where choices are made indiscriminately from the dictionary, and deletions similarly affect randomly selected words. As a result, the their synthetic machine translations $\widetilde{mt}$ may incorporate words that machine translation systems are unlikely to generate, leading to a qualitative deviation from the expected output of such systems.

3. Approach

3.1. Overview and Motivation

It is apparent from the existing methods that BT-BG and RAND have unique advantages. BT-BG is prone to generating translation errors that are likely to be found in (src, pe) pairs by learning of $(src, pe) \rightarrow mt$. On the other hand, RAND effectively limits the quantity of errors to align with the distribution found in gold data. In response to these observations, we introduce a noising method by adapting the masked language model (MLM), termed **MLM noising**. This method learns $(src, pe) \rightarrow mt$ and leverages the mask-infilling technique of MLM to create synthetic machine translations $\widetilde{mt}$ by introducing plausible translation errors into the reference (ref) of a parallel corpus while maintaining the frequency of these introduced errors to align with the error distribution found in gold data. Consequently, it presents a hybrid method that combines the advantages of both the BT-BG and RAND methods, effectively balancing error realism with controlled error quantity.

In addition, the TRANS method is conspicuous in that it utilizes raw machine translations (mt) that are not synthesized, thus closely replicating the generation process of mt found in gold data if the error frequency in mt aligns with that of gold data. To capitalize on the distinct advantages of both the TRANS and MLM noising methods, we propose a corpus ensemble method called **selective corpus interleaving**. This approach merges two distinct synthetic datasets (e.g., the datasets created by TRANS and MLM noising) into a single, enhanced dataset. This integration is achieved by comparing corresponding samples from the two datasets and
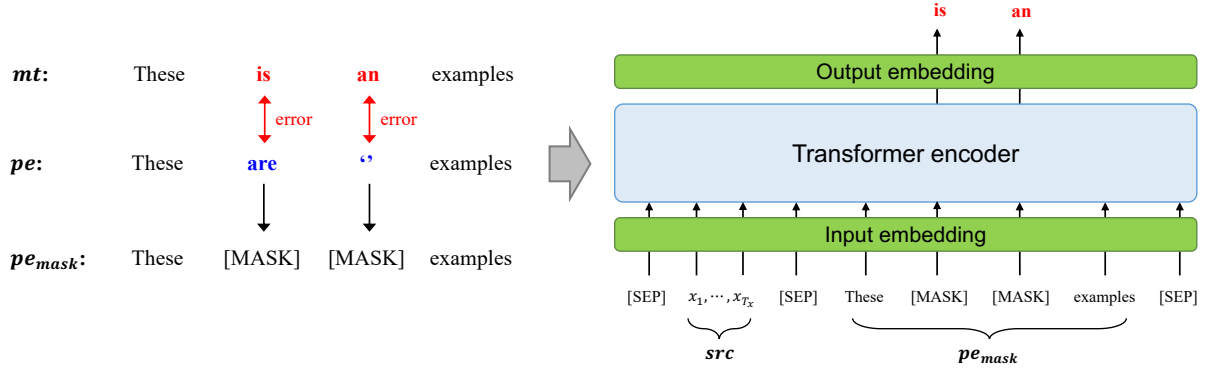
Figure 3: Overall architecture of our MLM noising model.

selectively incorporating those that more accurately reflect the attributes of gold data, thereby expecting to improve the overall quality and effectiveness of the synthetic data.

3.2. MLM Noising

3.2.1. Architecture and Objective

Given a corrupted text in which some tokens are replaced with `[MASK]`, the objective of MLM (Devlin et al., 2019) is to reconstruct the original text from this corrupted text by infilling masks with original tokens using the Transformer encoder (Vaswani et al., 2017). Inspired by studies (Kumar et al., 2020; Tuan et al., 2021) that employed MLMs for the data augmentation, we adapt MLM to produce a synthetic mt (i.e., $\widetilde{mt}$) that serves as a proxy for mt in our synthetic APE triplet.

As shown in Figure 3, given an APE triplet (src, mt, pe), we mask pe tokens that are aligned to **mistranslated tokens** in mt, after which we feed src along with this masked pe (pe_{mask}) and let the model learn to restore these **erroneous** mt tokens from the `[MASK]` tokens. In other words, this training objective indicates that the model learns to predict translation errors that may occur in (src, pe) pairs, leading to the learning of $(src, pe) \rightarrow mt$ that BT-BG likewise learns. We also note that our MLM noising can be regarded as a **cross-lingual MLM** (Conneau and Lample, 2019) because the two bilingual inputs, src and pe, constitute a cross-lingual representation by which the prediction is made on `[MASK]`. Formally, given that $\widehat{\mathbf{y}}$ denotes pe_{mask}, the objective function can be defined as

$$
\begin{aligned}
\max_{\theta} \ \mathcal{J}(\theta) &= P_\theta(\widetilde{\mathbf{y}} \mid \mathbf{x}, \widehat{\mathbf{y}}) \\
&\approx \sum_i m_i \log P_\theta(\widetilde{y}_i \mid \mathbf{x}, \widehat{\mathbf{y}}) \\
&= \sum_i m_i \log \frac{\exp\left(H_\theta(\widehat{y}_i)E_\theta^\top(\widetilde{y}_i)\right)}{\sum_{y'} \exp\left(H_\theta(\widehat{y}_i)E_\theta^\top(y')\right)},
\end{aligned}
\tag{2}
$$

where $H_\theta(\cdot)$ and $E_\theta(\cdot)$ are the hidden vector from Transformer and the embedding vector, respectively; and $m_i = 1$ indicates that $\widehat{y}_i$ is masked to correspond with mistranslated $\widetilde{y}_i$.

3.2.2. Training Data Configuration

For the learning process, it is imperative to construct a set of triplets in the form of (src, pe_{mask}, mt). At first, the alignment between mt and pe is necessary to construct pe_{mask}, but it is generally invisible in data. As an alternative, we utilize an **alignment**[2] that can be obtained as a byproduct of the edit distance[3] calculation between mt and pe. Given the edit-distance alignments from pe to mt denoted as $a = \{(y, \widetilde{y})_i\}_{i=1}^{T_{\widetilde{y}}}$, each alignment corresponds to one of four edit operations[4]: *keep, insert, delete, or substitute*. Specifically, for each $(y, \widetilde{y}) \in a$, the keep operation, where $y = \widetilde{y}$, represents that $\widetilde{y}$ is an accurate translation. Conversely, an alignment where $y \neq \widetilde{y}$ indicates a mistranslation, which is further categorized into (1) an insertion error if y is NULL, indicating the addition of unnecessary token; (2) a deletion error[5] if $\widetilde{y}$ is NULL, indicating the

---

[2] We obtain the alignments by using tercom software: https://github.com/jhclark/tercom.git

[3] The rationale for employing edit-distance alignment stems from the premise that pe results from only minor correction to mt, with the possibility that the majority of words between them remain unchanged. Consequently, we assume that discrepancies in word alignment between mt and pe—where aligned words are not identical—indicate that the word from mt is likely an erroneous translation.

[4] Due to implementation difficulties with mask-infilling, the shift operation employed in TER is not directly considered. Nonetheless, it is addressed indirectly, as it is widely recognized that a shift can be accomplished through a sequence of a deletion followed by an addition.

[5] In practice, we ignore alignments on deletion errors because all deletion `[MASK]` are mapped onto a single output token (i.e., null token ''), consequently causing serious deletion-biased predictions. Instead, we simulate deletion errors at the inference time.

487

Algorithm 1: ref_{mask} for training

Input: $\mathbf{y}, \widetilde{\mathbf{y}}$ # ref and mt
  D # distribution of error rate from gold data
  $a = \{(y, \widetilde{y})_i\}_{i=1}^{T_{\widetilde{y}}}$ # alignments
Output: $\widehat{\mathbf{y}}$
$\widehat{\mathbf{y}} \leftarrow \{\}$
$e \leftarrow \mathrm{Edit\_distance}(\widetilde{\mathbf{y}}, \mathbf{y})$ # actual error rate
$\hat{e} \sim Categorical(D)$ # sampled error rate
$a^{err} \leftarrow \{(y, \widetilde{y}) \mid \forall a, \widetilde{y} \neq \texttt{NULL}, y \neq \widetilde{y}\}$
if $e > \hat{e}$ **then**
  num_mask $\leftarrow \lceil(\mathrm{Len}(\mathbf{y}) * \hat{e})\rceil$
  $\hat{a} \leftarrow \mathrm{random\_choice}(a^{err}, \mathrm{num\_mask})$
else
  $\hat{a} \leftarrow a^{err}$
for each $(y, \widetilde{y}) \in a$ **do**
  if $\langle y, \widetilde{y}\rangle \in \hat{a}$ **then**
    $\widehat{\mathbf{y}} \leftarrow \mathrm{Append}(\widehat{\mathbf{y}}, \texttt{[MASK]})$
  else
    $\widehat{\mathbf{y}} \leftarrow \mathrm{Append}(\widehat{\mathbf{y}}, y)$

Algorithm 2: ref_{mask} for inference

Input: $\mathbf{y}, \mu = \{\mu_{\mathsf{keep}}, \mu_{\mathsf{ins}}, \mu_{\mathsf{del}}, \mu_{\mathsf{sub}}\}$
Output: $\widehat{\mathbf{y}}$
$\widehat{\mathbf{y}} \leftarrow \{\}$
op $\sim Categorical(\mu)$
for each $y \in \widehat{\mathbf{y}}$ **do**
  if op is keep **then**
    $\widehat{\mathbf{y}} \leftarrow \mathrm{Append}(\widehat{\mathbf{y}}, y)$
  else if op is insert **then**
    $\widehat{\mathbf{y}} \leftarrow \mathrm{Append}(\widehat{\mathbf{y}}, y)$
    $\widehat{\mathbf{y}} \leftarrow \mathrm{Append}(\widehat{\mathbf{y}}, \texttt{[MASK]})$
  else if op is delete **then**
    continue
  else if op is substitute **then**
    $\widehat{\mathbf{y}} \leftarrow \mathrm{Append}(\widehat{\mathbf{y}}, \texttt{[MASK]})$

3.2.3. Inference: Synthetic APE Data Construction

Once the training has finished, we use our MLM model to produce $\widetilde{mt}$ from the parallel corpus, yielding a new set of synthetic APE triplets $(src, \widetilde{mt}, ref)$. For every (src, ref) pair, we initiate the masking process for ref, wherein each token is masked **stochastically**[6] according to the categorical probability distribution[7] of the edit operations $\mu = \{\mu_{\mathsf{keep}}, \mu_{\mathsf{ins}}, \mu_{\mathsf{del}}, \mu_{\mathsf{sub}}\}$ obtained from gold data, facilitates diverse outcomes resulting from the informed masking decisions. We present this stochastic masking process to obtain ref_{mask} in Algorithm 2. Subsequently, we provide our MLM model with (src, ref_{mask}) and let it perform mask infilling to produce $\widetilde{mt}$.

Eventually, $\widetilde{mt}$ can hold the number of errors similar to gold data while these errors are expected to appear convincing by learning. Moreover, because our MLM stochastically determines masking patterns of ref_{mask} at every inference time, we can obtain $\widetilde{mt}$ in various forms from a single (src, ref) pair, making a given resource scalable.

3.3. Selective Corpus Interleaving

Recall that although TRANS is advantageous in using non-synthesized mt (which is most likely to capture the nature of mt in gold data), it has a limitation in including excessive errors on mt to match ref. We thus hypothesize that the benefit of synthetic mt that approximates the error statistics of gold data far outweighs the benefit of raw mt. However, it is intuitively reasonable that a TRANS triplet containing

omission of necessary token; and (3) a substitution error when both y and $\widetilde{y}$ are present but differ, indicating incorrect translation. Finally, we construct pe_{mask} by replacing pe tokens, which are aligned with mistranslated mt tokens, with [MASK] tokens.

Due to the limited availability of gold data, we transform synthetic data from the TRANS triplets (src, mt, ref) into $(src, ref_{mask}, \widetilde{mt})$ for our training needs. Notably, TRANS typically exhibits a higher error rate compared to gold data, necessitating the use of $\widetilde{mt}$ to approximate the edit distance between mt and pe found in gold data.

To align the error distribution of synthetic data with that of gold data, we first calculate the proportion of masking tokens based on the error rates observed in the gold data (by sampling from the categorical distribution in Figure 2). For each ref in the TRANS triplet, we sample the number of tokens (k) to be masked and find the total tokens (n) that are aligned with mistranslated mt tokens based on edit-distance alignment. Next, we randomly choose k tokens out of the n tokens, replace these with their aligned mt tokens to form $\widetilde{mt}$, and mask them to create ref_{mask}. This procedure yields a set $\mathcal{S} = \{(\widehat{y}, \widetilde{y}_i)\}_{i=1}^{|\mathcal{Y}|}$, where $|\mathcal{Y}| = \binom{n}{k}$ represents the total number of potential masking scenarios, each leading to distinct pairs of $(ref_{mask}, \widetilde{mt})$. The model is trained to recognize these masking patterns across iterations, aiming for generating appropriate errors at various positions by

$$\max_{\theta} \; \mathbb{E}_{(\widehat{\mathbf{y}}, \widetilde{\mathbf{y}}) \sim \mathcal{S}} \left[\sum_i m_i \log P_\theta(\widetilde{y}_i \mid \mathbf{x}, \widehat{\mathbf{y}}) \right]. \quad (3)$$

We present our pseudo algorithm for constructing ref_{mask} from the TRANS triplet (src, mt, ref) in Algorithm 1.

---

[6]Note that the alignment used in the training phase is intrinsically unavailable during inference.

[7]Each μ represents the probability that the corresponding edit operation occurs. This distribution can be obtained from the edit-distance calculation: https://github.com/jhclark/tercom.git
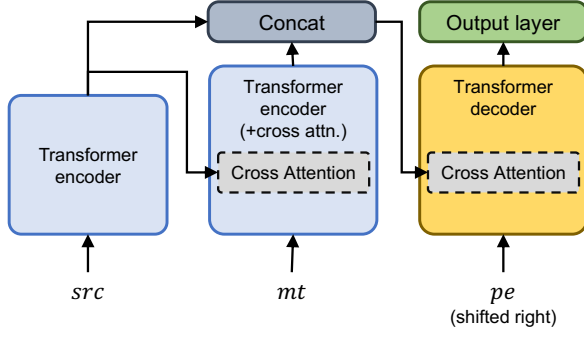
Figure 4: Schematic architecture of the concat-based APE model, one of the basic APE models that performs well.

mt whose error quantity is already similar to that of gold data can be regarded as an ideal training sample for APE, which can be even more effective than using $\widetilde{mt}$ instead (e.g., the overlapping region in Figure 2).

In this regard, we suggest a selective corpus interleaving approach that takes only advantageous samples between TRANS and ours depending on whether the mt–ref edit distance is similar to the mt–pe edit distance, thereby constructing a single enhanced synthetic dataset. For every (src, ref) pair, we select either mt or $\widetilde{mt}$ by applying the three-sigma rule (Pukelsheim, 1994):

$$ mt = \begin{cases} mt & \text{if } |\text{edit}(mt, ref) - \mu| \leq \lambda\sigma \\ \widetilde{mt} & \text{otherwise,} \end{cases} \quad (4) $$

where $\text{edit}(\cdot)$ denotes the edit distance; μ and σ are the mean and standard deviation of the mt–pe edit distances, respectively; and $\lambda \in [1, 3]$ is a hyperparameter.

4. Experiments

4.1. Setup

Evaluation Metric Following the WMT APE shared task (Chatterjee et al., 2018), we adopted TER (↓) (Snover et al., 2006)[8] as our primary metric, and BLEU (↑) (Koehn et al., 2007)[9] as the secondary metric. We conducted all evaluations case-sensitively.

Datasets We used the WMT'18 APE dataset as the gold data, a human-made English-German (EN-DE) APE dataset considered the *de-facto* standard benchmark. The WMT data consists of 23K of training data, 1K of development data, and three separate test datasets (Test16, Test17, and Test18)

of 2K each. We also used EN-DE TRANS synthetic triplets[10] consisting of approximately 7M samples. All words in the datasets we used were tokenized into subword units by SentencePiece[11].

Model Configuration Our MLM noising model was developed utilizing the RoBERTa architecture (Liu et al., 2019)[12], adhering largely to its default configuration. Specifically, the model comprises 12 layers, each with 12 attention heads, a hidden layer size of 768, and a feed-forward layer size of 3,072. For training, we employed the AdamW optimizer (Loshchilov and Hutter, 2019), configured with β parameters of (0.9, 0.999), a learning rate of 2e-4 subject to linear decay, 7,000 warm-up steps, and managed a batch size of 384 samples.

For the implementation of an APE model, we utilized OpenNMT-py[13], specifically deploying the "concat-based" model (Figure 4) that aligns with state-of-the-art standards. In line with established practices, the model architecture was configured with 6 layers, 8 attention heads, hidden layer sizes of 512, and feed-forward network sizes of 2,048. Training was conducted using the Adam optimizer (Kingma and Ba, 2015) with β set to (0.9, 0.998), following the learning rate schedule recommended by Vaswani et al. (2017), which includes 6,000 warm-up steps. The model was trained with a batch size capable of accommodating 48K tokens.

4.2. Experimental Details

In our experiment, we establish our baselines using four models trained on the existing synthetic datasets discussed in Section 2: TRANS, BT-FG[14], BT-BG[14], and RAND. To ensure a fair and controlled comparison, each model was trained under identical conditions, including hyperparameters, codebase, and training seed, with the sole exception of the synthetic training data. Furthermore, all synthetic datasets, including our own, were created using the same parallel corpus used to generate TRANS so that they only differed in the mt portion.

Since our MLM noising and RAND can diversify their $\widetilde{mt}$ as stochastic, we empirically used five different random seeds to construct them; namely, one out of five different $\widetilde{mt}$ was selected for each (src, ref) pair at every iteration during training. The training of the MLM model lasted 3 days with 8 A100 GPUs and the APE model lasted for 12 hours with a single A5000 GPU.

---

[8] https://github.com/jhclark/tercom
[9] https://github.com/moses-smt/mosesdecoder

[10] https://ict.fbk.eu/escape/
[11] https://github.com/google/sentencepiece
[12] https://huggingface.co/roberta-base
[13] https://github.com/OpenNMT/OpenNMT-py
[14] https://github.com/wonkeelee/APE-backtranslation.git

	Test16		Test17		Test18		Test Avg.	
Approach	TER($\downarrow$)	BLEU($\uparrow$)	TER($\downarrow$)	BLEU($\uparrow$)	TER($\downarrow$)	BLEU($\uparrow$)	TER($\downarrow$)	BLEU($\uparrow$)
Trans	16.87	73.95	17.30	73.08	17.80	72.41	17.32	73.15
BT-FG	17.26	73.56	17.56	72.78	17.89	72.14	17.57	72.82
BT-BG	17.61	73.04	17.60	72.49	18.01	71.89	17.74	72.47
Rand	17.23	73.59	17.61	72.69	17.81	72.38	17.55	72.88
MLM Noising (w/o interleave)	16.90	74.03	17.31	72.90	17.62	72.43	17.28	73.12
MLM Noising (w/ interleave)	**16.71**	**74.58**	**16.74**	**73.79**	**17.43**	**72.88**	**16.96**	**73.75**

Table 1: Comparison of learning effects on each synthetic data by presenting evaluation results of the model trained on each specific synthetic data. In each column, the best-performing result is highlighted in **bold**.

	Test Avg.		Sample Ratio	
	TER($\downarrow$)	BLEU($\uparrow$)	MLM	Trans
$\lambda = 0$ (MLM Noising)	17.32	73.15	100.0%	0.0%
$\lambda = 1$	17.25	73.26	71.5%	28.5%
$\lambda = 2$	**16.96**[**]	**73.75**[**]	41.5%	58.5%
$\lambda = 3$	17.03[**]	73.48[**]	20.8%	79.2%
$\lambda = \infty$ (Trans)	17.28	73.12	0.0%	100.0%

Table 2: Effect of the selective corpus interleaving with varying λ. ** indicates the improvement is statistically significant compared to both $\lambda = 0$ and $\lambda = \infty$ with $p < 0.01$. The **bold** highlights indicate the best result in each column.

4.3. Results

We evaluated the APE models, each trained on distinct synthetic datasets, by employing the WMT test datasets. This allowed us to investigate how each synthetic dataset impacts their performance. Referring to Table 1, our observations yield two key findings: (1) The utilization of synthetic data generated through MLM noising improves the APE performance compared to the baselines, and (2) augmenting this with selective corpus interleaving method further enhances the model performance. In light of these results, we speculate that our findings lend support to our initial hypothesis, suggesting potential benefits for APE training. Specifically, the combination of MLM noising, drawing from the strengths of BT-BG and Rand, and interleaving MLM noising with Trans, where the former focuses on statistics while the latter aligns more closely with gold data, demonstrates promise in improving the APE performance.

Additionally, we performed experiments employing various values of λ for the corpus interleaving method (Table 2). When $\lambda = 0$, it signifies the exclusive use of the dataset generated by MLM noising, while $\lambda = \infty$ indicates the exclusive use of the Trans dataset. Notably, our results revealed that taking approximately equal parts ($\lambda = 2$) from our MLM noising and Trans data led to better APE performance than the other ratios. These findings also support our hypothesis that each synthetic dataset
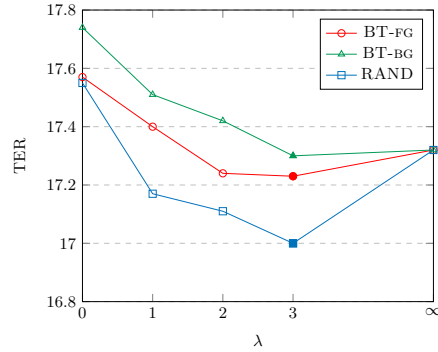

Figure 5: Effect of the selective corpus interleaving when applied to other existing synthetic APE data. The colored plots represent the best performance for each data.
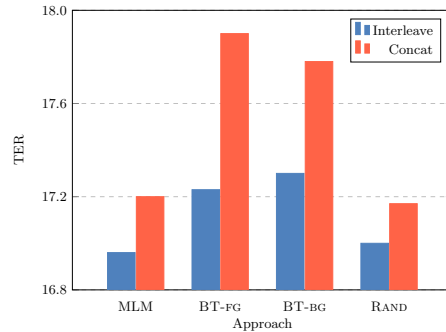

Figure 6: Comparison of two corpus ensemble methods applied to each synthetic APE dataset.

possesses unique advantages that complement one another.

5. Analysis and Discussion

5.1. Effect of Selective Corpus Interleaving

The versatility of the selective corpus interleaving technique allows for its application across a variety of pre-existing synthetic datasets, including BT-FG,

Component	Sentence
src	It is worthless without knowing where to sell it .
ref	Zu wissen , wo man es verkauft , ist sehr wichtig .
$\widetilde{mt}$ (MLM noising)	Zu wissen , wo man es **zu** verkauft , **wo sie** sehr wichtig .
mt (Trans)	**Es ist nichts ohne zu** wissen , wo **sie zu verkaufen** .

(a)

Component	Sentence
src	What happens if I want to leave ?
ref	Was geschieht , wenn ich wieder gehen will ?
$\widetilde{mt}$ (MLM noising)	Was **passiert** geschieht , wenn ich **ich zu** gehen **wollen .**
mt (Trans)	Was **passiert** , wenn ich **verlassen wollen** ?

(b)

Table 3: Two examples showing the comparison between mt and $\widetilde{mt}$ for a given (src, ref) pair. The **bold** highlights indicate the mistranslated words compared to ref.

BT-BG, and RAND, provided that these datasets feature synthetic machine translations $\widetilde{mt}$ aimed at reducing the edit distance to ref. To investigate this utility further, we undertook additional experiments where selective corpus interleaving was employed to integrate TRANS data with each of the aforementioned datasets (BT-FG, BT-BG, and RAND). The outcomes, illustrated in Figure 5, indicate that utilization of the selective corpus interleaving with these alternate synthetic datasets consistently improves the quality of the synthetic dataset, thereby contributing to enhanced performance of the models.

Moreover, we conducted a comparative analysis contrasting the selective corpus interleaving method with a conventional corpus ensemble technique that straightforwardly concatenates two datasets without any selective filtering. The outcomes (Figure 6) validate that the selective corpus interleaving approach significantly surpasses the straightforward concatenation method in performance, even though the latter offers a twofold increase in data resources. This finding highlights the importance of excluding filtered samples, an crucial step in the effective application of the selective corpus interleaving method, underscoring its impact on improving data quality.

5.2. Case Study

In Table 3, we present two examples to examine the differences between our modified translation output by MLM noising, denoted as $\widetilde{mt}$, and the raw translation output, referred to as mt (representing TRANS). The first example, shown in Table 3a, illustrates that mt requires extensive post-processing to align with the reference (ref). In contrast, $\widetilde{mt}$, generated through our approach, necessitates fewer corrections, showing a closer resemblance to the

error statistics found in gold data. However, in our second example, presented in Table 3b, the number of errors in mt is already within an acceptable range, making it more compatible with the gold standard. In such cases, the benefits of using $\widetilde{mt}$ are less pronounced, and the selective corpus interleaving method ultimately favors mt.

Additionally, our observations have identified several potential weakness in our approach:

- Due to the nature of the masked language modeling, all masked tokens are reconstructed independently of one another. This independence can occasionally lead to undesirable words, such as repeated words especially when consecutive positions in ref are masked during the inference process. Nevertheless, this independence may be acceptable as long as it doesn't occur excessively, as our primary goal is to introduce errors intentionally.

- Unlike insert and substitute operations, delete operations are applied randomly to construct the masked reference (ref_{mask}) for MLM noising. While some level of randomness is acceptable for our error introduction purposes, it can occasionally alter the semantics of the text (e.g., "New York City" becoming "New City").

- The edit-distance alignment used to create the masked post-edited text (pe_{mask}) for MLM training does not always correspond to semantic errors. Consequently, $\widetilde{mt}$ may sometimes contain unnatural or incorrect words.

6. Related Work

A similar approach involving the use of MLM for generating synthetic data was proposed by Tuan

491

et al. ([2021](#)) in the context of the quality estimation task that is similar to the APE task in terms of that models are trained on pairs of (src, mt) Their method shares some similarities with ours: (1) they employed MLM to create synthetic mt from a parallel corpus, and (2) for a given ref, they randomly replaced tokens with [MASK] or deleted tokens and inserted [MASK] tokens at random positions. Subsequently, they input src and the masked ref into their MLM model to perform mask infilling, resulting in synthetic mt.

Nevertheless, there are several notable distinctions between our approach and theirs: (1) In their masking procedure, the choice of tokens for insertion, deletion, and substitution was independent of the statistical properties of the gold data; (2) They employed a pre-trained, off-the-shelf multilingual BERT model ([Devlin et al., 2019](#)), which was trained on clean text, and thus, their MLM was incapable of inducing errors in mt from [MASK] tokens; (3) Despite addressing multilingual texts, their multilingual BERT model did not incorporate cross-lingual features, thereby limiting its ability to establish a joint representation between src and mt.

7. Conclusion

In this study, we introduce a method for adapting MLM to generate synthetic APE data from a parallel corpus. This method is further refined through the incorporation of selective corpus interleaving. We summarize our key findings as follows:

- We present a training approach for MLM, designed to predict translation errors from masked tokens. By utilizing this trained MLM model, we generate synthetic mt data by introducing errors into ref through a process known as mask infilling.

- Additionally, we propose the selective corpus interleaving method, a technique for effectively merging two distinct datasets. This method involves the selection of samples whose error profiles closely resemble those found in the gold data, bridging the gap between the datasets.

- Our experimental results and analysis highlight the importance of ensuring that synthetic data closely mirrors the error statistics of gold data.

We believe that our work offers opportunities for further extensions and improvements by incorporating different MLM techniques. For instance, approaches like ELECTRA ([Clark et al., 2020](#)) that employ adversarial learning within MLM can help determine the plausibility of mask prediction results, resulting in more natural and realistic synthetic data.

8. Limitations

Our method, while promising, presents several challenges that need to be addressed in future:

- The edit distance-based alignments utilized in our training may not always correspond to semantic discrepancies, potentially resulting in the generation of translation errors during inference that fail to reflect actual translation inaccuracies. This discrepancy might result in the production of errors that appear unnatural.

- The inherent nature of MLM, which treats each masked token independently, makes them prone to generating plausible yet contextually inappropriate tokens. Consequently, token generation from our MLM noising process may not always meet the nuanced requirements of natural language generation.

- Although the MLM model's error distribution aims to mimic that of gold APE data (i.e., benchmark datasets), it is crucial to recognize that these gold datasets do not capture every possible translation error. As a result, the MLM noising approach may inherently bias towards the types of translation errors prevalent in the benchmark datasets used for training, potentially restricting the diversity of errors it generates.

9. Bibliographical References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

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.

BSI. 1973a. *Natural Fibre Twines*, 3rd edition. British Standards Institution, London. BS 2570.

BSI. 1973b. Natural fibre twines. BS 2570, British Standards Institution, London. 3rd. edn.

A. Castor and L. E. Pollux. 1992. The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Rajen Chatterjee, Markus Freitag, Matteo Negri, and Marco Turchi. 2020. Findings of the WMT 2020 Shared Task on Automatic Post-Editing. In *Proceedings of the Fifth Conference on Machine Translation*, pages 646–659, Online. Association for Computational Linguistics.

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

Rajen Chatterjee, Marion Weller, Matteo Negri, and Marco Turchi. 2015. Exploring the planet of the APEs: a comparative study of state-of-the-art methods for MT automatic post-editing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 156–161, Beijing, China. Association for Computational Linguistics.

J.L. Chercheur. 1994. *Case-Based Reasoning*, 2nd edition. Morgan Kaufman Publishers, San Mateo, CA.

N. Chomsky. 1973. Conditions on transformations. In *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. `arXiv:2003.10555 [cs.CL]`.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Umberto Eco. 1990. *The Limits of Interpretation*. Indian University Press.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Paul Gerhard Hoel. 1971a. *Elementary Statistics*, 3rd edition. Wiley series in probability and mathematical statistics. Wiley, New York, Chichester. ISBN 0 471 40300.

Paul Gerhard Hoel. 1971b. *Elementary Statistics*, 3rd edition, Wiley series in probability and mathematical statistics, pages 19–33. Wiley, New York, Chichester. ISBN 0 471 40300.

Otto Jespersen. 1922. *Language: Its Nature, Development, and Origin*. Allen and Unwin.

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.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages

177–180, Prague, Czech Republic. Association for Computational Linguistics.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26.

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, 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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. `arXiv:1907.11692 [cs.CL]`.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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).

Friedrich Pukelsheim. 1994. The Three Sigma Rule. *The American Statistician*, 48(2):88–91.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Jaehun Shin, Wonkee Lee, Byung-Hyun Go, Baikjin Jung, Youngkil Kim, and Jong-Hyeok Lee. 2021. Exploration of effective attention strategies for neural automatic post-editing with transformer. 20(6).

Charles Joseph Singer, E. J. Holmyard, and A. R. Hall, editors. 1954–58. *A history of technology*. Oxford University Press, London. 5 vol.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, Cambridge, Massachusetts, USA.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).

S. Superman, B. Batman, C. Catwoman, and S. Spiderman. 2000. *Superheroes experiences with books*, 20th edition. The Phantom Editors Associates, Gotham City.

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, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6000–6010, Long Beach, California, USA. Curran Associates Inc.

Wanrong Zhu, Zhiting Hu, and Eric Xing. 2019. Text Infilling. `arXiv:1908.08566 [cs.CL]`.