Enhancing Large Language Models for Document-Level Translation Post-Editing Using Monolingual Data

Zongyao Li*, **Zhiqiang Rao***, **Hengchao Shang***,
Jiaxin GUO, Shaojun Li, Daimeng Wei, Hao Yang
Huawei Translation Service Center, Beijing, China
{lizongyao,raozhiqiang,shanghengchao,
guojiaxin1,lishaojun18,weidaimeng,yanghao30}@huawei.com

Abstract

The translation capabilities of neural machine translation (NMT) models based on the encoder-decoder framework are extremely potent. Although Large Language Models (LLMs) have achieved remarkable results in many tasks, they have not reached state-of-the-art performance in NMT. However, traditional NMT still faces significant challenges in areas of document translation such as context consistency, tense, and pronoun resolution, where LLMs inherently possess substantial advantages. Instead of directly using LLMs for translation, employing them for Automatic Post-Editing (APE) to post-edit NMT outputs proves to be a viable option. However, document-level bilingual data is extremely scarce. This paper proposes a method that can effectively leverage the capabilities of LLMs to optimize document translation using only monolingual data. By employing two NMT models in opposite directions (Source-to-Target and Target-to-Source), we generate pseudo-document training data for the training of APE. We have identified and resolved the issue between training and inference mode inconsistency brought about by the pseudo-document training data. The final experimental results demonstrate that by using only document-level monolingual data, we can significantly improve the quality of NMT and greatly enhance issues such as reference and contextual consistency in NMT.

1 Introduction

Large Language Models (LLMs) exhibit outstanding performance in a multitude of tasks (Touvron et al., 2023; Anil et al., 2023; Bahak et al., 2023). Despite this, in various domains, sentence-level Neural Machine Translation (NMT) models are not inferior to LLMs in metrics such as BLEU (Papineni et al., 2002) and COMET (Rei et al., 2022), but they still face many challenges in document-level translation, especially in terms of contextual consistency, tense, and reference. Given that LLMs are primarily trained using document-level monolingual data, they undoubtedly have a significant advantage in terms of the coherence of textual expression.

Recently, a great deal of research (Wang et al., 2023; Zhang et al., 2023a; Guo et al., 2024) has been devoted to the use of LLM to enhance the quality of MT. The research primarily uses LLM directly for translation and focuses on addressing the following issues: 1) LLMs are all centered on English and need to align cross-lingually (Zhang et al., 2023b; Alves et al., 2023) to improve the translation ability of LLMs. Current methods mainly rely on high-quality bilingual alignment data for pre-training or supervised fine-tuning (SFT) to achieve this, but the final results do not show significant advantages over sentence-level NMT models in terms of BLEU and COMET metrics. 2) It is necessary to investigate the quality of LLM in multilingual (Liu et al., 2024) and domain-specific NMT scenarios (Zheng et al., 2024), especially in domain-specific scenarios where traditional NMT models have already performed well. None of the above methods has integrated the traditional NMT model with LLMs, failing to take advantage of the respective strengths of both.

Koneru et al. (2024) uses LLMs to perform automatic post-editing (APE) on NMT-generated translations, rather than directly using LLMs for translation. First, bilingual document-level data are used with NMT models to construct (SRC, NMT→PE) triplet training data. Second, these training data are used to fine-tune the LLMs, enabling them to refine the translations produced by NMT models. Finally, the fine-tuned LLM is used to optimize the outputs of the NMT model, improving the consistency and coherence of the translations. However, the limitation of the method described in the paper is the scarcity of bilingual data at document level, especially in low-resource scenarios, which limits

*These authors contributed equally to this work.

Proceedings of the 31st International Conference on Computational Linguistics, pages 8830–8840
January 19–24, 2025.

the applicability of the method to multilingual and low-resource settings.

Compared with obtaining bilingual data at the document-level, it is much easier to obtain monolingual data. In Voita et al. (2019), the DocRepair method was proposed, which only requires monolingual data at the document-level in the target language. It is a sequence-to-sequence model that maps inconsistent sentence groups to consistent ones. The consistent group comes from the original document-level monolingual data; the inconsistent group is obtained by sampling from back-and-forth translations of each isolated sentence. The method proposed can also improve the consistency of sentence-level NMT translations well even when using only monolingual data. However, it is a cascaded system, which may amplify errors; in addition, there is no reference to the source text, and the information such as entity consistency in the source text is not effectively utilized.

We proposes a method for optimizing document translation by leveraging the capabilities of LLMs using only monolingual document-level data. It involves fine-tuning LLMs to APE via LoRA (Hu et al., 2021) for optimizing sentence-level NMT system outputs. Similar to DocRepair, by translating the target-side text back and forth twice, we obtain mapping data for inconsistent and consistent sentence groups. Unlike DocRepair, we also include the generated source MT data in the mapping group, forming a triple data (SRC, MT→PE). Since no bilingual document-level data is used, the SRC in our triple data is pseudo MT data, while during inference, SRC is natural data, causing issues of mismatch between training and inference patterns that impact translation quality. Wei et al. (2023) mentions significant stylistic differences between MT and natural data. To address this mismatch issue, we conduct style transfer on the pseudo SRC data to bridge the gap with natural data. By jointly using monolingual document-level data from source and target, we enable LLMs to refine the sentence-level NMT system outputs and generate consistent and coherent text by utilizing their ability to produce fluent and lengthy documents.

The main findings and contributions of this article are as follows:

1. We propose a new paradigm that enables LLMs to have document-level machine translation APE capability using only document-level monolingual data. Especially in low-resource scenarios, the quality of sentence-level MT models can be improved by leveraging low-cost monolingual data. Simultaneously, we demonstrate that by increasing the quantity of document-level monolingual data, the translation quality can reach, or even surpass, the effects of using limited document-level bilingual data.

2. We have proposed various methods (Cascade, End2End) of using LLMs for style transfer to address the issue of mismatch between training and inference, and experimental results have demonstrated the effectiveness of our method.

3. We analyzed the translation after APE, and achieved an accuracy of 80.2% on the ContraPro English to German test set, indicating that our method can significantly improve entity consistency.

2 Method

We use the Large Language Model (LLM) as an Automatic Post-Editing (APE) tool to improve the translation quality of sentence-level Neural Machine Translation (NMT) models. We only use document-level monolingual data, in conjunction with two sentence-level NMT models, to construct pseudo document-level training data through one round of back-translation and one round of forward-translation, for fine-tuning the LLM. Since there is a difference between the pseudo data and real data, during the training phase, we use the LLM to bridge this gap. The inference process of the translation system based on APE is a cascaded process: 1) Obtain the translation using the sentence-level NMT model. 2) Post-edit the translation using the APE model.

2.1 Using LLM as APE

APE is a sub-field of machine translation that automatically corrects the errors that reside in the machine translation results. The training data for the APE model consists of triplet data: source sentence (SRC), machine translation result (MT), and post-edited sentence (PE). During training, the APE can be divided into two modes based on whether SRC is used.

$$L(\theta) = \arg\max_{\theta} \log P(PE|MT;\theta) \quad (1)$$

$$L(\theta) = \arg\max_{\theta} \log P(PE|(SRC,MT);\theta) \quad (2)$$

where θ is the parameter of the APE model.

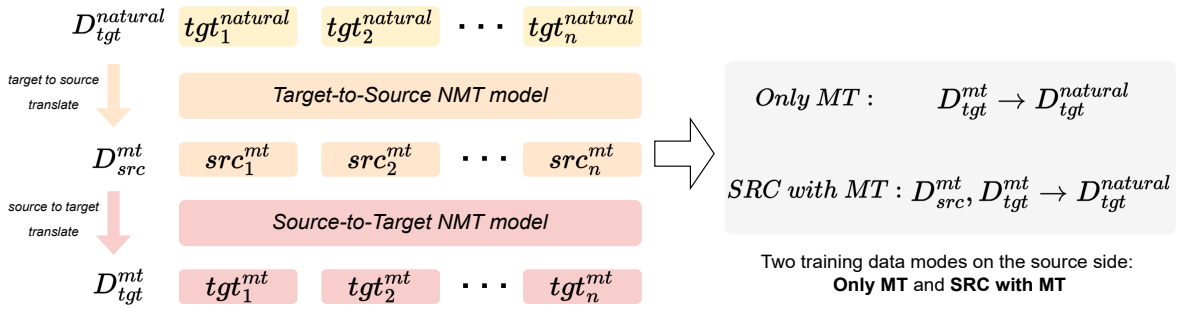

Figure 1: On the left is the process of constructing APE training data with monolingual using NMT models, and on the right are the two data modes required for training APE.

The advantage of formula 1 is its ability to construct training data without the need for bilingual data, by leveraging a backward NMT model and monolingual data from the target side. However, this approach may amplify errors in machine translation. In contrast, formula 2 can simultaneously utilize both SRC and MT information, but it requires aligned bilingual data.

2.2 Pseudo-Document-Level Bilingual Data Generation

As shown in the figure 1, we construct the pseudo document-level bilingual data through the following steps:

1. Choose monolingual corpora in the target language and extract a sequence of n consecutive sentences $D_{tgt}^{natural} = \{tgt_1^{natural}, tgt_2^{natural}, ..., tgt_n^{natural}\}$ from it.

2. For all sentences from 1 to n, translate them using the target-to-source MT model to obtain $D_{src}^{mt} = \{src_1^{mt}, src_2^{mt}, ..., src_n^{mt}\}$. Then translate the translated D_{src}^{mt} using the source-to-target MT model to obtain $D_{tgt}^{mt} = \{tgt_1^{mt}, tgt_2^{mt}, ..., tgt_n^{mt}\}$. The decoding of NMT models are all based on beam search, and the beam size is 5.

3. For APE training, the D_{src}^{mt} and D_{tgt}^{mt} generated by the NMT model are SRC and MT, while monolingual data $D_{tgt}^{natural}$ corresponds to PE. The training data form corresponding to formula 1 is **Only MT** mode: $D_{tgt}^{mt} \rightarrow D_{tgt}^{natural}$, while that corresponding to formula 2 is **SRC with MT** mode: $D_{src}^{mt}, D_{tgt}^{mt} \rightarrow D_{tgt}^{natural}$.

2.3 Bridge the Gap between Training and Inference

Intuitively, **SRC with MT** mode should perform better as it leverages the original text information. In fact, our APE model is also based on **SRC with MT** mode. However, the subsequent experimental results show that directly training on the pseudo data using **SRC with MT** mode results in negative gains. The main reason for this is the significant difference between the SRC in the training data, which is D_{src}^{mt}, and the SRC during inference, which is $D_{src}^{natural}$. To bridge this gap, we propose two methods (to express this more clearly, we provide an example using a translation task where the source language is English (en) and the target language is German (de)). Figure 2 details our two training methods:

Cascade method: First, we take English as the target language and construct the training data $D_{en}^{mt} \rightarrow D_{en}^{natural}$ with **Only MT** mode, and using the formula 1 to train LLM as APE_{en}.

Then, taking German as the target language, we construct the training data $D_{en}^{mt}, D_{de}^{mt} \rightarrow D_{de}^{natural}$ with **SRC with MT** mode, use APE_{en} to post edit on D_{en}^{mt} and get $D_{en}^{natural'}$, obtain the final training data $D_{en}^{natural'}, D_{de}^{mt} \rightarrow D_{de}^{natural}$, and fine-tune on LLM with the final training data to obtain the final APE model.

End2End method: We use English as the target language and construct the training data $CPT_{en} = \{D_{en}^{mt}, D_{en}^{natural}\}$, using the **Only MT** mode. Meanwhile, we use German as the target language and construct the $CPT_{de} = \{\{D_{de}^{mt}, D_{de}^{natural}\}, \{D_{en}^{mt}, D_{de}^{mt}, D_{de}^{natural}\}\}$, using both modes.

For the training data during the Continual Pre-training stage (CPT), we introduce the newline tag between the MT sentences and the natural language sentences. We first pre-train the LLM as LLM_{pre} with these data. The data CPT_{en} is intended for LLM to align the stylistic differences between the original text side MT and natural; the data CPT_{de} aim to equip LLM with certain target language APE capabilities during pre-training. During pre-

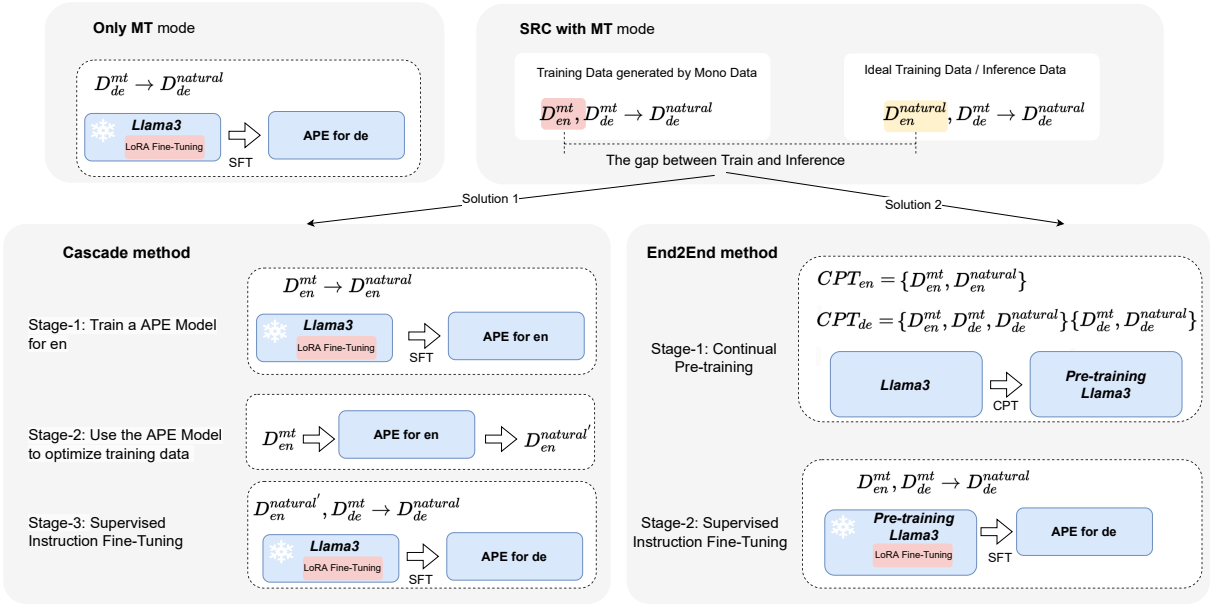

Figure 2: At the top of this figure are two modes for applying APE to optimize MT using LLM: **Only MT** mode and **SRC with MT** mode. Below are two proposed methods to address the gap between training and inference in the **SRC with MT** mode, which arises from using only monolingual constructed training data. On the left is the Cascade method, where the LLM first optimizes the source text before model training. On the right is the End2End method, which incorporates two different types of data (MT and natural data) during the pre-training phase, enabling the LLM to learn the underlying relationships between the two and subsequently achieve better results in the SFT stage.

training, we utilize a word dropout (=0.1) (Sennrich et al., 2016) strategy to ensure training convergence.

Afterwards, by taking German as the target language, we use **SRC with MT** mode to construct training data $D_{en}^{mt}, D_{de}^{mt} \rightarrow D_{de}^{natural}$. We directly apply the constructed training data to fine-tune the pre-trained LLM_{pre} with supervised fine-tuning (SFT), resulting in an APE model.

2.4 Inference

Since document translation involves context, its decoding is relatively complex, and there are currently several decoding methods: 1) **Chunk-Based**, which is a simple method that divides the document to be translated into approximately equal-sized non-overlapping chunks and translates them independently; 2) **Batched Sliding Window**, proposed by Post and Junczys-Dowmunt (2024), translate the document by appending the sentence to be translated with preceding context, using a sliding window approach with a payload, and extract the last sentence after translating the entire chunk; 3) **Sequential Decoding**, proposed by Herold and Ney (2023), append the left source context and use forced decoding of the previous sentence's translation as the target context in the next step, translating only one sentence at each step. The results of Koneru et al. (2024) indicate that the effects of these three decoding methods are similar, so in this paper, we only use the Chunk-Based decoding method, where the size of the chunk is related to the maximum sequence length during inference.

3 Experimental Setup

Model: We used two sentence-level NMT models, source-to-target and target-to-source, to assist in generating training data. The architecture of the sentence-level NMT model is transformer-big, with the encoder and decoder consisting of 25 and 6 layers, respectively. It is first trained on the English-German data from WMT2023[1], and then fine-tuned using the MuST-C V3 corpus(Di Gangi et al., 2019). For LLM, we conducted experiments using the latest open-source Llama3-8B[2] model by Meta. During training, we masked the loss on the prompt, meaning that the LLM was trained exclusively to predict the reference based on the given source and hypothesis. We utilized the transformers library for training and inference with Llama3. When training the adapters, we con-

[1]https://www2.statmt.org/wmt23/translation-task.html
[2]https://llama.meta.com/llama3/

figured the hyperparameters with rank 8, alpha 32, dropout 0.1, and set the bias as 'LoRA_only'. Following the approach by Dettmers et al. (2023) to enhance the model's robustness to LoRA hyperparameters, we adapted on all layers. The added modules to the adapter include q_proj, k_proj, v_proj, $gate\_proj$, up_proj, and $down\_proj$. During training and inference, we set the max tokens to 1024.

Dataset: We use the MuST-C V3 corpus as the main training and test data; additionally, to verify whether larger-scale document-level monolingual data can further improve the results, we select 100,000 documents from the MC4 dataset (Xue et al., 2021).

Evaluation Metrics We use COMET-22 (COMET) (Rei et al., 2022) as the primary evaluation metric, and additionally use sentence-level BLEU (Papineni et al., 2002) and document-level SacreBLEU (Liu et al., 2020) to evaluate translation quality, denoted as s-bleu and d-bleu, respectively.

Consistency Evaluation We also report the scores of ContraPro (Müller et al., 2018) to assess context-specific resolution of pronoun ambiguity. We constructed the test set based on the methods proposed in Koneru et al. (2024) and the MuDA tagger (Fernandes et al., 2023), and evaluated the three metrics: Pronouns, Formality, and Lexical Cohesion.

4 Main Results

To validate the effectiveness of the method we proposed, we selected a translation task from English to German on the MUST-C V3 dataset. We assess the following configurations:

Sen2Sen: This is a sentence-level NMT model. Its generation process involves firstly training on the English-German dataset provided by WMT2023, then fine-tuning on MuST-C V3. This is a strong baseline model, and the quality of the generated translations is comparable to some commercial NMT models.

Llama3 Few shot APE: We construct several prompts to guide Llama3 to perform post-edit on MT. All our prompts refer to the format of Koneru et al. (2024).

Llama3 MT Sen2Sen: We process the dataset of MuST-C V3 into training data aligned at sentence level and use Llama3 for LoRA fine-tuning.

Llama3 MT Doc2Doc: We process the MuST-C V3 dataset into aligned document data of approximately equal length using the Chunk-Based method, and perform Lora fine-tuning using Llama3.

APE for para-Doc: We use the document-level bilingual data in MuST-C V3 to replicate the method proposed in Koneru et al. (2024) based on Llama3 and source-to-target NMT model.

APE for Cascade: Based on the Cascade method, we first use the English data from MuST-C V3 to train APE_{en}, then use the German data from MuST-C V3 to construct pseudo-document data, and use APE_{en} to perform PE on the corresponding data $D^{mt}_{en}, D^{mt}_{de} \rightarrow D^{natural}_{de}$, finally obtain the model trained by formula 2.

APE for End2End: Based on the End2End method, we first separately use the en and de from MuST-C V3 as target languages to generate training data for **Only MT**, pre-train on Llama3. Then, we use the pseudo-document data generated with de as the target language from **SRC with MT** and obtain the model trained with formula 2.

APE for Large Mono: First, a larger volume of Mono data (100,000) is extracted from the MC4 dataset, and then training is performed based on the **APE for End2End** method.

4.1 Automatic Post-Editing is Necessary.

Given that Llama3 has seen significant improvements over Llama2 (Touvron et al., 2023) in both training methods and data volume, we need to re-evaluate whether APE for LLMs is still effective. As shown in Table 1, we first use a prompt to guide Llama3 to perform post-editing on MT, which is the result of Llama3 Few-shot APE. It is evident that Llama3 indeed possesses strong post-editing capabilities, capable of enhancing translation quality. Additionally, the APE approach outperforms the direct use of LLMs for translation, such as Llama3 MT Doc2Doc and Llama3 MT Sen2Sen, with improvements of more than 0.5 points in s-bleu, d-bleu, and COMET metrics. The conclusion that "Automatic Post-Editing is Necessary," as obtained by Koneru et al. (2024) on Llama2, has been further validated on Llama3.

4.2 The Monolingual Document-Level Data is Effective.

As shown in Table 1, in the case of using only monolingual data, the Cascade and End2End methods we proposed can approach the results obtained by (Koneru et al., 2024) using actual bilingual data.

System	s-bleu	d-bleu	COMET	Pronouns			Formality			Lexical Cohesion		
				P	R	F1	P	R	F1	P	R	F1
Sen2Sen	34.63	38.78	84.68	0.67	0.77	0.72	0.70	0.72	0.71	0.62	0.75	0.68
Llama3 Few shot APE	34.70	38.84	84.85	0.67	0.78	0.72	0.71	0.73	0.71	0.62	0.76	0.68
Llama3 MT Sen2Sen	31.29	35.52	85.03	0.66	0.76	0.71	0.68	0.71	0.69	0.61	0.76	0.68
Llama3 MT Doc2Doc	34.30	38.60	85.62	0.69	0.81	0.75	0.72	0.80	0.76	0.61	0.77	0.68
APE for para-Doc	35.03	39.27	85.85	0.70	**0.83**	**0.76**	**0.75**	**0.82**	**0.78**	**0.63**	0.77	0.69
Our approach												
APE for Cascade	34.85	38.81	85.76	0.69	0.81	0.75	0.73	0.80	0.76	0.62	0.77	0.69
APE for End2End	34.74	38.71	85.71	0.69	0.80	0.74	0.72	0.80	0.76	0.62	0.76	0.68
APE for Large Mono	**35.48**	**39.60**	**86.50**	**0.71**	0.82	**0.76**	**0.75**	0.81	**0.78**	**0.63**	**0.78**	**0.70**

Table 1: Comparing various optimization translation methods with Llama3. For document decoding, we use chunk-based decoding. We report metrics such as s-bleu, d-bleu, COMET, MuDA tagger. The best score in each metric is highlighted in bold.

System	Contra(%)
Sen2Sen	66.7
APE for para-Doc	87.3
APE for Cascade	80.2
APE for End2End	79.5

Table 2: Comparing document APE accuracy on the ContraPro English→German test set.

System	s-bleu	d-bleu	COMET
Sen2Sen	34.63	38.78	84.68
APE for para-Doc	35.03	39.27	85.85
APE for **Only MT**	34.64	38.87	85.16
APE for **SRC with MT**	32.90	37.03	84.60
APE for Cascade	34.85	38.81	85.76
APE for End2End	34.74	38.71	85.71

Table 3: Comparing the impacts of using different forms of SRC in training data.

Compared to the Sen2Sen results, the NMT translations, after being refined with the LLM fine-tuned using our proposed methods for APE, show minor changes in s-bleu and d-bleu, but there is a 1-point increase in COMET, along with varying degrees of improvement in Pronouns, Formality, and Lexical Cohesion metrics. After introducing 100,000 document-level monolingual data, our methods (APE for Large Mono) can significantly outperform the results from bilingual data across various metrics. This demonstrates that using document-level monolingual data for APE is effective.

4.3 Enhancement of APE for Contextual Phenomena

In the table 1, the three indicators of Pronouns, Formality, and Lexical Cohesion are constructed in our test sets by referencing the methods proposed in Koneru et al. (2024) and MuDA tagger (Fernandes et al., 2023). We also select talks with the highest number of tags for each phenomenon, resulting in 14 talks in our test sets, addressing contextual occurrences related to pronouns, formality, and lexical cohesion. Then, we remove the selected talks from our training data and use them for testing. Experimental results show that the Cascade and End2End methods both enhance the discourse characteristics of NMT.

We further validate the enhancement of APE in disambiguating pronouns using the ContraPro test set, which is a benchmark specifically designed to assess the disambiguation of pronouns, namely "Er" (masculine), "Sie" (feminine), and "Es" (neutral) when translating "It" from English to German. We referenced the evaluation methods from Post and Junczys-Dowmunt (2024); Koneru et al. (2024). Table 2 indicates that, although our proposed method exhibits 7 differences compared to APE for para-Doc, both Cascade and End2End show an improvement of over 13 points on the Sen2Sen basis.

The above results indicate that LLM is very suitable for dealing with tasks involving contextual information and plays a very prominent role in contextual phenomena.

5 Analysis

5.1 Bridge the Gap between Training and Inference

In Section 2.3, the issue of training and inference mismatch was mentioned, and several sets of ex-

Training Data	s-bleu	d-bleu	COMET
baseline (Sen2Sen)	34.63	38.78	84.68
APE for End2End	34.74	38.71	85.71
random sample	34.48	38.54	84.71
domain sample	34.62	38.67	85.42

Table 4: The results of conducting experiments on different datasets using the End2End method.

periments were designed to illustrate and address this issue. As shown in Table 3, **APE for Only MT** and **APE for SRC with MT** were obtained by generating data using **Only MT** and **SRC with MT** respectively in Section 2.2, and then training APE. APE for **Only MT** was able to improve COMET by 0.5; however, with the addition of SRC, D_{src}^{mt}, the results of APE for **SRC with MT** saw a decrease of more than 1 point in both s-bleu and d-bleu, and there was also a slight decrease in COMET. This indicates that directly using D_{src}^{mt} does not bring positive gains to APE.

APE for **SRC with MT** and APE for para-Doc have exactly the same training process, but yield significantly different results, indicating a substantial gap between the D_{src}^{mt} used during training and the $D_{src}^{natural}$ used during inference. In our following experiments, we compared and analyzed the differences between pseudo source text and actual source text. We trained a binary classifier for D_{src}^{mt} and $D_{src}^{natural}$ using fast-text, and found that 87% of the data could be distinguished. However, after applying the Cascade method to D_{src}^{mt}, only 58% of the data could be correctly classified. The results of APE for Cascade show that compared to APE for **SRC with MT**, s-bleu, d-bleu, and COMET all show an improvement of more than 1 point, indicating that our proposed Cascade method can bridge the gap between D_{src}^{mt} and $D_{src}^{natural}$ effectively.

The problem of APE for Cascade lies in the fact that using LLM to generate training data is a costly and inelegant method. We believe that LLM can implicitly possess the matching capabilities of D_{src}^{mt} and $D_{src}^{natural}$. The results of APE for End2End indicate that, through training in the pre-training phase, LLM can address this gap, although there is still a very small difference compared to Cascade.

5.2 Has Data Leak Resulted in Positive Outcomes?

In the End2End method, we use both the Source and Target of the MuST-C V3 data as monolingual data. Could this introduce a bilingual data leakage issue during the pre-training stage? We set up the following experiments to verify this:

1) **random sample**: In order to maintain consistency with the dataset of MuST-C, 2500 English and 2500 German documents were randomly extracted from the mC4 dataset.

2) **domain sample**: fast-text[3] was used to classify MC4 by domain, and 2500 English and 2500 German documents close to the domain of MuST-C were selected respectively.

Then, the method of APE for End2End is used to experiment on the two data selection methods mentioned above. As shown in table 4, the improvement of random sample is relatively small, while the results of domain sample can basically approach the results obtained using the MuST-C data (APE for End2End). This indicates that there is no issue of data leaking, and the quality improvement is largely related to domain adaptation in LLM.

5.3 The Impact of Training Data Scale

We further experimented on the impact of different amounts of monolingual data on APE quality. The data amount of MUST-C V3 is 2537 documents. We also used the method of domain sample from MC4 to select 2500, 5000, 10000, 20000, 40000 and 100000 documents for experiments of different data scale. The table 5 indicate that with the increase of data volume, the quality of APE can indeed be improved. When using 20000 documents, the quality of the model has already surpassed the result of para-Doc. Further increasing the data amount resulted in a diminishing increase in quality, and there was hardly any significant improvement in the results beyond 40000 documents.

5.4 The Adaptability of APE

We want to know if the APE system trained with pseudo-corpus constructed by a NMT system can be adapted to other NMT systems. We divided the training data of NMT into two parts and trained two NMT models: nmt_1 and nmt_2. We then used these two NMT models with the End2End method to train APE separately, obtaining ape_1 and ape_2. Table 6 shows that different NMT systems can achieve better adaptation only when trained with their own NMT-generated pseudo-corpus; when nmt_1 is combined with ape_2 and nmt_2 is combined with ape_1, the improvement in COMET's performance is somewhat reduced. This indicates

[3]https://github.com/facebookresearch/fastText

Training data scale	2500	5000	10000	20000	40000	100000	para-Doc
COMET	85.42	85.64	85.76	86.16	86.42	86.50	85.85

Table 5: Comparing the impact of using different scales of domain sample data.

NMT	APE	COMET
nmt_1		84.52
nmt_2		84.55
nmt_1	ape_1	85.61
nmt_1	ape_2	85.23
nmt_2	ape_1	85.30
nmt_2	ape_2	85.58

Table 6: The COMET results of pairing two NMT and APEs for pairwise combinations.

that our APE system has a certain generality, but optimal results can only be achieved when used in conjunction with the NMT used to construct the training data.

6 Related Work

In the era of training NMT models with sequence-to-sequence patterns, many methods have already been proposed to optimize document translation. Starting from Tiedemann and Scherrer (2017), consecutive sentences in the context have been used to assist the translation of the current sentence. Agrawal et al. (2018); Zhang et al. (2022) and others have also demonstrated that translating multiple sentences together can improve translation quality. Additionally, many new methods for model architecture such as HAN (Miculicich et al., 2018), SAN (Maruf et al., 2019), HYBRID (Zheng et al., 2020), FLATTRANS (Ma et al., 2020), GTRANS (Bao et al., 2021) have been proposed to better encode document information; due to the scarcity of aligned document data, data augmentation methods like MULTIRRES (Sun et al., 2022), IADA (Wu et al., 2024b) have also been proposed to more efficiently utilize limited aligned data; furthermore, the DocRepair (Voita et al., 2019) method can use document-level monolingual data, combined with the encoder-decoder mode of the APE model, to improve translation consistency; Feng et al. (2022) use cache technology to cache global information for enhancing the local information of the current sentence; and the DocRerank method proposed by Yu et al. (2020) reorders the results of sentence-level models in the decoding stage, selecting sentences or tokens with high document consistency.

The LLM has achieved very impressive results in a large number of natural language processing tasks. At the same time, more and more people are conducting research on large models in the field of machine translation. Xu et al. (2024); Guo et al. (2024) mainly focus on sentence-level machine translation research, focusing on how to strengthen the cross-lingual alignment capabilities of large models. There are also some work (Ding et al., 2024; Gao et al., 2023) to introduce Retrieval Augmented Generation (RAG), coupled with translation memories (Mu et al., 2023; Moslem et al., 2023) or to correct NMT system output in prompts (Raunak et al., 2023; Chen et al., 2024). Wu et al. (2024a); Wang et al. (2023) meticulously explored the enhancement of document-level translation using large models; Koneru et al. (2024) proposed using LLM for APE instead of directly for translation, which also yielded good results.

7 Conclusion

We introduced a new method to enhance document translation quality by leveraging only monolingual data and large language models. This approach fine-tunes LLMs on Automatic Post-Editing using novel strategies to improve the output of sentence-level neural machine translation. We identified and resolved the issue of inconsistency between training and inference modes caused by the use of synthetic training data. This method significantly improves translation quality and consistency, as validated by a notable score on the ContraPro test set, making it particularly useful in settings where bilingual data is scarce.

8 Limitations

Our method, a cascaded system, involves initial decoding with NMT followed by LLM decoding, leading to higher inference latency and error magnification. It is ideal for low-resource scenarios, particularly in specialized fields like biomedicine, where bilingual data is limited. Future research will focus on this area.

References

Ruchit Agrawal, Marco Turchi, and Matteo Negri. 2018. Contextual handling in neural machine translation: Look behind, ahead and on both sides. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 31–40, Alicante, Spain.

Duarte M. Alves, Nuno M. Guerreiro, João Alves, José Pombal, Ricardo Rei, José G. C. de Souza, Pierre Colombo, and André F. T. Martins. 2023. Steering large language models for machine translation with finetuning and in-context learning. *Preprint*, arXiv:2310.13448.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report. *Preprint*, arXiv:2305.10403.

Hossein Bahak, Farzaneh Taheri, Zahra Zojaji, and Arefeh Kazemi. 2023. Evaluating chatgpt as a question answering system: A comprehensive analysis and comparison with existing models. *Preprint*, arXiv:2312.07592.

Guangsheng Bao, Yue Zhang, Zhiyang Teng, Boxing Chen, and Weihua Luo. 2021. G-transformer for document-level machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3442–3455, Online. Association for Computational Linguistics.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2024. Iterative translation refinement with large language models. *Preprint*, arXiv:2306.03856.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. 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 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Yujuan Ding, Wenqi Fan, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meets llms: Towards retrieval-augmented large language models. *Preprint*, arXiv:2405.06211.

Yukun Feng, Feng Li, Ziang Song, Boyuan Zheng, and Philipp Koehn. 2022. Learn to remember: Transformer with recurrent memory for document-level machine translation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1409–1420, Seattle, United States. Association for Computational Linguistics.

Patrick Fernandes, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig. 2023. When does translation require context? a data-driven, multilingual exploration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 606–626, Toronto, Canada. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Jiaxin Guo, Hao Yang, Zongyao Li, Daimeng Wei, Hengchao Shang, and Xiaoyu Chen. 2024. A novel paradigm boosting translation capabilities of large language models. *Preprint*, arXiv:2403.11430.

Christian Herold and Hermann Ney. 2023. Improving long context document-level machine translation. In *Proceedings of the 4th Workshop on Computational Approaches to Discourse (CODI 2023)*, pages 112–125, Toronto, Canada. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Sai Koneru, Miriam Exel, Matthias Huck, and Jan Niehues. 2024. Contextual refinement of translations: Large language models for sentence and document-level post-editing. *Preprint*, arXiv:2310.14855.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2024. Is translation all you need? a study on solving multilingual tasks with large language models. *Preprint*, arXiv:2403.10258.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Preprint*, arXiv:2001.08210.

Shuming Ma, Dongdong Zhang, and Ming Zhou. 2020. A simple and effective unified encoder for document-level machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3505–3511, Online. Association for Computational Linguistics.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3092–3102, Minneapolis, Minnesota. Association for Computational Linguistics.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

Yongyu Mu, Abudurexiti Reheman, Zhiquan Cao, Yuchun Fan, Bei Li, Yinqiao Li, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. 2023. Augmenting large language model translators via translation memories. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10287–10299, Toronto, Canada. Association for Computational Linguistics.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post and Marcin Junczys-Dowmunt. 2024. Escaping the sentence-level paradigm in machine translation. *Preprint*, arXiv:2304.12959.

Vikas Raunak, Amr Sharaf, Yiren Wang, Hany Hassan Awadallah, and Arul Menezes. 2023. Leveraging gpt-4 for automatic translation post-editing. *Preprint*, arXiv:2305.14878.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. Rethinking document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3537–3548, Dublin, Ireland. Association for Computational Linguistics.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten,

Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. Context-aware monolingual repair for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 877–886, Hong Kong, China. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

Daimeng Wei, Zhanglin Wu, Hengchao Shang, Zongyao Li, Minghan Wang, Jiaxin Guo, Xiaoyu Chen, Zhengzhe Yu, and Hao Yang. 2023. Text style transfer back-translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7944–7959, Toronto, Canada. Association for Computational Linguistics.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024a. Adapting large language models for document-level machine translation. *Preprint*, arXiv:2401.06468.

Minghao Wu, Yufei Wang, George Foster, Lizhen Qu, and Gholamreza Haffari. 2024b. Importance-aware data augmentation for document-level neural machine translation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 740–752, St. Julian's, Malta. Association for Computational Linguistics.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. *Preprint*, arXiv:2309.11674.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. *Preprint*, arXiv:2010.11934.

Lei Yu, Laurent Sartran, Wojciech Stokowiec, Wang Ling, Lingpeng Kong, Phil Blunsom, and Chris Dyer. 2020. Better document-level machine translation with bayes' rule. *Preprint*, arXiv:1910.00553.

Biao Zhang, Ankur Bapna, Melvin Johnson, Ali Dabirmoghaddam, Naveen Arivazhagan, and Orhan Firat. 2022. Multilingual document-level translation enables zero-shot transfer from sentences to documents. *Preprint*, arXiv:2109.10341.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. *Preprint*, arXiv:2301.07069.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023b. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481, Singapore. Association for Computational Linguistics.

Jiawei Zheng, Hanghai Hong, Xiaoli Wang, Jingsong Su, Yonggui Liang, and Shikai Wu. 2024. Fine-tuning large language models for domain-specific machine translation. *Preprint*, arXiv:2402.15061.

Zaixiang Zheng, Xiang Yue, Shujian Huang, Jiajun Chen, and Alexandra Birch. 2020. Towards making the most of context in neural machine translation. *Preprint*, arXiv:2002.07982.