Post-Editing with Error Annotation for Machine Translation: Dataset Construction using GPT-4

Youyuan Lin[1] Masaaki Nagata[2] Chenhui Chu[1]
[1]Kyoto University [2]NTT
youyuan@nlp.ist.i.kyoto-u.ac.jp masaaki.nagata@ntt.com chu@i.kyoto-u.ac.jp

Abstract

Despite advancements in Large Language Models (LLMs) and their state-of-the-art Machine Translation (MT) performance, fine-grained Automatic Post Editing (APE) remains useful for MT by aiding in error detection, especially when applied to specific domains. To train such APE systems that incorporate error detection capabilities, APE datasets with error annotation are essential. However, most APE datasets are lack of error annotations. We aim to construct error-annotated APE datasets that improve the error-based editing proficiency of APE systems. We leverage GPT-4, expand post-editing into error-annotated MT datasets. By employing a Chain-of-Thought (CoT) setting that applies an error annotation-based analysis step before post-editing, we ensure the post-editing corrects the errors while maintaining the phraseology. Our experiments on a Japanese-English medical MT dataset reveal that GPT-4, coupled with CoT and error annotations, excels in generating high-quality, annotation-informed post-edits.

1 Introduction

Machine Translation (MT) models, particularly those trained on web-derived parallel corpora, are reported to suffer from reliability issues despite their high average performance, especially when meeting the specific lexicon and stylistic requirements of certain application domains [1]. Automatic Post-Editing (APE) is a post-processing task in a Machine Translation workflow, aiming to automatically identify and correct errors in MT outputs [1]. An APE system is usually acquired by training on datasets comprising triplets of three texts: source (*src*), machine-translation (*mt*) of *src*, and post-edited sentence (*pe*) of *mt*.

With the remarkable strides made in MT systems, particularly the evolution of Large Language Models (LLMs) in the past two years, we have witnessed a dramatic enhancement in the capabilities of state-of-the-art MT systems [2]. This technological leap brings to the forefront a compelling query: Why not solely rely on these more advanced MT systems for high-quality translations instead of engaging in post-editing?

In response to this question, we contend that the value of post-editing transcends mere improvement in translation quality. A fine-grained APE encompasses the critical functions of detecting, elucidating and rectifying errors within the translations. The post-editing generated by an APE system should based on the error annotations. This process not only refines the output but also contributes to the ongoing improvement and understanding of MT systems.

To achieve such a fine-grained APE process, we argue that the APE dataset needs to come forth with accurate error annotations (*err*). On this issue, Multidimensional Quality Metrics (MQM) [3] provides the span and category for errors as the error annotations. It is considered as the most reliable human evaluation framework for MT [4]. There have been many datasets annotated based on MQM or MQM-like framework [4, 5]. Then, the challenge lies in using MQM-annotated (*src*, *mt*, *err*) triplets to generate accurate post-edited translations (*pe*), requiring not just error correction but a nuanced understanding of the errors to maintain the integrity and style of the source while enhancing translation quality.

In this work, we explore the efficacy of GPT-4 on the task of construction of APE datasets with error annotations in a Chain-of-Thought (CoT) setting, i.e., applying an error analysis step to the translations before post-editing. We evaluate our method APE on a Japanese-to-English medical MT dataset and extend this dataset with fine-grained post-editings. We show that GPT-4 produces meaningful human-judgment-aligned *pe* that also leads to general

quality improvements. Also, we find GPT-4 is capable of accurately using specialized terminology based on reference. The results demonstrate that CoT with error annotation is critical in constraining the pe to be close to the initial translation while fixing the errors.

2 The Automatic Post-Editing Task

In most settings (e.g., the WMT task on MT Automatic Post-Editing,[1)]) the post-editing task is formalized as follows: given src and mt, generate the most likely pe, i.e.:

$$pe = \underset{\hat{pe}}{\operatorname{argmax}}\, P(\hat{pe}|src, mt) \quad (1)$$

In which $P(\hat{pe}|src, mt)$ denotes the probability of a post-editing $\hat{pe}$ given src and mt. Compared to the vanilla MT task, the distinction of the APE task is that the mt is also entered as the input. It is conceivable that, in extreme cases, a powerful enough system could generate accurate pe based only on the src without referring to the mt. In such case, the new translation may not qualify as an "edited text" but is simply a "better quality translation." The whole task remains a vanilla MT task.

3 Method

Instead, we want the APE system to be based on the mt to correct errors while retaining the phraseology as much as possible. In other words, we believe that compared to the MT system, the APE system, while outputting better quality translations, should also correct errors in initial translations while preserving other correct expressions. Thus, the APE system should identify errors in the translation before editing the translation. Then, the task should be formalized as follows:

$$pe, err = \underset{\hat{pe},\hat{err}}{\operatorname{argmax}}\, P(\hat{pe}, \hat{err}|src, mt) \quad (2)$$

During the APE process, the APE system is tasked with concurrently identifying errors in the translated text. The objective is to preserve the original diction and grammatical structure, correcting only the identified errors.

To train a system accomplishes the above task, we need to include datasets with quaternions (src, mt, err, pe). Many studies have shown that LLMs can perform a series of complex tasks through In-Context-Learning (ICL) [6] and CoT [7]. We propose to constructing the APE dataset with the above quaternions according to LLMs.

1) http://www2.statmt.org/wmt23/ape-task.html

Given a MT dataset annotated under MQM framework[3], comprising triplet (src, mt, err), we generate and expand pe into this dataset. For ICL, multiple demonstrations are provided, each using (src, mt, err) as input. We meticulously craft error-annotation based pe outputs for these demonstrations, with the quality of pe being validated by native speakers. In our prompts, we implement CoT, structuring the model's task into two phases: initially generating a detailed description of err, followed by editing the mt based on this description. This two-step approach ensures a thorough understanding and accurate correction of errors, enhancing the overall quality and reliability of the post-edited output.

APE systems are frequently employed in domain-specific environments [1] to rectify inaccuracies generated by MT systems trained on general corpora. In such specialized contexts, maintaining the terminological precision of the constructed dataset is paramount. To make the terminology output by an MT model accurate, a common approach is to fine-tune the MT model for domain adaptation [8]. It usually takes extra time and computation resources. During the dataset construction phase, we enhance terminological accuracy by utilizing human-annotated references (ref) when available. Note that ref may not serve as a qualified pe to the mt, due to its lack of direct reference to the mt. Specifically, We use the ref also as an input, provide a task description in the propmt and instruct the model to refer to the terminology that occurs in the ref. Demonstrations are provided to show how the same terminology as ref is used in pe without replicate the ref.

4 Experimental Settings

Dataset: We experiment with an English-to-Japanese Medical MT dataset (EJMMT) published by Arase et al. [5], which is annotated with error spans and error types. The annotation identified 4, 492 errors on 1, 903 translation output, each with at least one error. Table 1 shows a few samples from the dataset, covering all the error types. All errors fall into a customized error typology containing the following five types:

- Addition: The target text includes text not present in the source.
- Omission: Content is missing from the translation that is present in the source.
- Mistranslation: The target content does not accurately

src	*mt*	*ref*
Regular exercise can improve both of these qualities.	通常の運動は、これらの性質を改善することができる。	定期的な運動によってその両方を向上させることができます。
Even former athletes who stop exercising do not retain measurable long-term benefits.	運動をやめた元スポーツ選手でさえ、長期的な長期的な利益を維持することはできない。	元運動選手であっても、運動をやめてしまえば、その効果を長期間維持することはできません。
Endoscopic cyanoacrylate injection: Doctors pass an endoscope through the mouth and into the digestive tract.	内視鏡的シアノアクリラート注射: 医師は内視鏡を通して内視鏡を通して消化管に入ります。	内視鏡的シアノアクリレート注入：内視鏡を口から消化管に挿入します。

Table 1 Selected examples in the EJMMT dataset. Colored texts indicate the annotated error spans, and the type corresponds to the following: Addition, Omission, Mistranslation, Grammar, Terminology.

represent the source content.

- Grammar: Syntax or function words are presented incorrectly.
- Terminology: The target text is not suitable in terms of the domain of the document.

Note that 1,697 errors fall in "Terminology" catagory inside 4,492 errors. MT translations are output half by Google's neural MT system [9] and half by NICT's neural MT system [10]. For each source sentence, a corresponding Japanese translation is available as the reference, prepared by human translators with a professional review.

To compare the methods under various settings, we randomly selected 64 samples from the dataset as the test set and did experiments on them. We then used the optimal setting on the experiments for the entire dataset.

Prompt: We experiment with gpt (gpt-4-turbo) in our experiments. We experiment under four prompt settings:

- (i) gpt: We use a prompt that describes the system's role as a translator.
- (ii) gpt + *mt*: We use a prompt that describes the system's role as a post-editor.
- (iii) gpt + *mt* + CoT: We use a prompt that describes the system's role as a post-editor and under the CoT setting.
- (iv) gpt + *mt* + CoT + *ref*: We use a prompt that describes the system's role as a post-editor, and under the CoT setting, the system can refer to the *ref* for terminologies.

For each setting, we provide 8 demonstrations covering all the error types in the EJMMT dataset. All used prompts are shown in appendix A.1.

Metrics and Evaluation: For quality evaluation, we use two reference-free Quality Estimation (QE) models: COMET-22 (wmt-22-cometkiwi-da) [11] and COMET-23 (wmt-23-cometkiwi-da-xl) [12]. Higher scores represent higher quality. We use COMET-* to denote both COMET-22 and COMET-23 below. To measure the similarity with *mt*, we use the average Translation Edit Rate (TER) [13] implementation from PyTER[2]. Texts with lower TER need fewer edits to make the text the same as the *mt*, indicating that the text is closer to *mt* regarding phraseology.

	COMET-22	COMET-23	TER
mt	83.67	70.93	-
ref	85.83	75.45	54.43
gpt	86.61	77.27	47.63
gpt + *mt*	86.46	76.86	45.95
gpt + *mt* + CoT	86.74	77.47	**38.38**
gpt + *mt* + CoT + *ref*	**86.78**	**77.67**	38.76

Table 2 Results on the randomly selected samples (N = 64). A situation with a higher BLEU score but a lower TER with *mt* indicates a better result. Bold items stand for the optimal results.

	COMET-22	COMET-23	TER
mt	83.87	71.69	-
ref	84.95	74.68	56.04
gpt	**86.72**	**77.50**	44.67
gpt + *mt* + CoT + *ref*	86.33	77.16	**39.32**

Table 3 Results on the whole dataset (N = 1,903). A situation with a higher BLEU score but a lower TER with *mt* indicates a better result. Bold items stand for the optimal results.

5 Results

Translation Quality improvement Table 5 shows the results on the randomly selected 64 samples. Throughout, we find that: First, all settings' COMET-* scores exceeded the *ref* human translators gave, showing that the edited text has a higher translation quality with these settings. Meanwhile, their TERs are also lower than the *ref*, which means that the words and phrases of the edited text are stylistically closer to the machine-translated text. These indicators provide evidence for the high quality of the dataset constructed based on GPT-4. Second, comparing setting

2) https://github.com/roy-ht/pyter

Source	*mt*	*ref*	gpt	gpt + *mt* + CoT + *ref*
These beneficial effects in turn decrease the risk of heart attack, stroke, and coronary artery disease.	これらの有益な効果は、心臓発作、脳卒中、および冠状動脈疾患のリスクを減少させる。	こうした有益な効果によって、心臓発作、脳卒中、**冠動脈疾患**のリスクが低下します。	これらの有益な効果は、その結果、心臓発作、脳卒中、および冠状動脈疾患のリスクを減少させます。	これらの有益な効果により、順に心臓発作、脳卒中、および**冠動脈疾患**のリスクを減少させます。
An extra or abnormal chromosome may increase the risk of cancer.	余分な染色体または異常な染色体があると、がんのリスクを高めることになります。	**染色体の過剰**や異常ががんのリスクを増大させることもあります。	余分なまたは異常な染色体は、がんのリスクを高める可能性があります。	**過剰な染色体**または異常な染色体があると、がんのリスクが増大する可能性があります。

Table 4 Selected examples of the post-editings. Bold texts denote a accurate medical terminology. Colored text denote an inaccurate medical terminology.

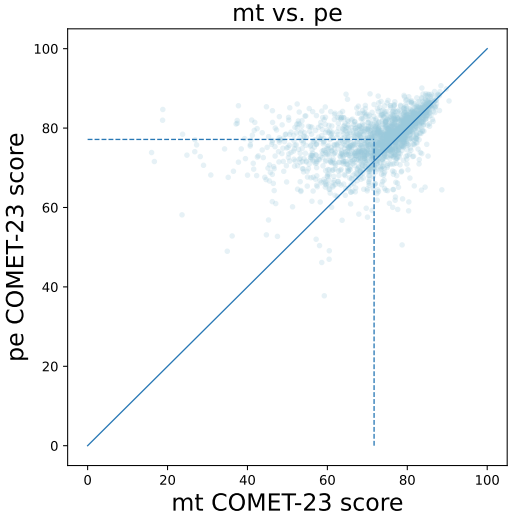

Figure 1 The COMET-23 scores before and after editing the entire dataset. The X-axis is the score of *mt*, and the Y-axis is the score of *pe*. *pe* stands for the gpt + *mt* + CoT + *ref* setting. Each point denotes a sentence in the dataset. The solid line represents the line $y = x$. The dotted line indicates the average score. 74.35% of the sentences get improved in score after editing.

(ii) with setting (i), the inclusion of *mt* did not increase the COMET-* score; it even lowered it. TER is slightly lower. Both settings behave similarly. This result suggests that the mere inclusion of *mt* allows GPT-4 to treat the task as MT rather than APE. Third, according to the results of settings (iii) and (iv), CoT dramatically influences the way GPT-4 performs its tasks. The TER was considerably reduced. At the same time, the COMET-* score maintains the original GPT4 level. These results lead us to believe that GPT-4 can provide high-quality translated text, and CoT is an effective method to constrain GPT-4 to perform APE tasks rather than MT tasks.

We constructed the post-editing dataset using setting (iv) on the entire dataset. Table 5 shows the results. Similar to the results obtained in tests using small samples, the quality of GPT-4 translations exceeds that of the *ref* supplied by human experts. With CoT, the newly generated translations are closer to *mt* while improving quality. Figure 1 shows the COMET-23 scores before and after editing the entire dataset in detail. In this figure, 74.35% of the translations are improved after post-editing.

Accurate Use of Terminology

Table 5 presents a curated selection of examples illustrating this aspect. These instances reveal that while GPT-4 translations exhibit proficiency, they are not immune to occasional inaccuracies in terminological choices. However, the incorporation of *ref* enhances GPT-4's capability in picking terminology. It adeptly extracts and applies accurate medical terminology from *ref*, demonstrating its utility in bridging the gap between automated translations and the exactitude required in medical contexts.

6 Summary

This study has demonstrated the remarkable potential of leveraging LLMs for APE in domain-specific machine translation. Our experimental results convincingly show that GPT-4 can perform high-quality translation post-editing in the English-to-Japanese medical domain. This improvement is evident even when compared to the human-crafted reference translations. It's also shown that incorporating a CoT approach with error annotation proved crucial in steering GPT-4's capabilities toward effective APE. This methodology ensures that the post-edited translations adhere closely to the original texts regarding phraseology while simultaneously correcting errors.

We expand the EJMMT dataset with post-editings and error annotation in natural language. In the next step, we will try to train an APE model with the same high performance and smaller size using the expanded dataset.

Acknowledgement

This work was supported by NTT.

References

[1] Pushpak Bhattacharyya, Rajen Chatterjee, Markus Freitag, Diptesh Kanojia, Matteo Negri, and Marco Turchi. Findings of the wmt 2023 shared task on automatic post-editing. In **Proceedings of the Eighth Conference on Machine Translation**, pp. 672–681, 2023.

[2] Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. Multilingual machine translation with large language models: Empirical results and analysis. **arXiv preprint arXiv:2304.04675**, 2023.

[3] Valerie R Mariana. **The Multidimensional Quality Metric (MQM) framework: A new framework for translation quality assessment**. Brigham Young University, 2014.

[4] Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. Experts, errors, and context: A large-scale study of human evaluation for machine translation. **Transactions of the Association for Computational Linguistics**, Vol. 9, pp. 1460–1474, 2021.

[5] Takeshi Hayakawa and Yuki Arase. Fine-grained error analysis on english-to-japanese machine translation in the medical domain. In **Proceedings of the 22nd Annual Conference of the European Association for Machine Translation**, pp. 155–164, 2020.

[6] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey for in-context learning. **arXiv preprint arXiv:2301.00234**, 2022.

[7] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. **Advances in Neural Information Processing Systems**, Vol. 35, pp. 24824–24837, 2022.

[8] Chenhui Chu and Rui Wang. A survey of domain adaptation for neural machine translation. **arXiv preprint arXiv:1806.00258**, 2018.

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

[10] Rui Wang, Masao Utiyama, Andrew Finch, Lemao Liu, Kehai Chen, and Eiichiro Sumita. Sentence selection and weighting for neural machine translation domain adaptation. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, Vol. 26, No. 10, pp. 1727–1741, 2018.

[11] Ricardo Rei, Marcos Treviso, Nuno M Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José GC De Souza, Taisiya Glushkova, Duarte M Alves, Alon Lavie, et al. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. **arXiv preprint arXiv:2209.06243**, 2022.

[12] Ricardo Rei, Nuno M Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José GC de Souza, and André FT Martins. Scaling up cometkiwi: Unbabel-ist 2023 submission for the quality estimation shared task. **arXiv preprint arXiv:2309.11925**, 2023.

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

[14] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. **arXiv preprint arXiv:2309.03409**, 2023.

A Appendix

A.1 The Prompt

Following recent work on automatic prompt engineer [14], the prompt used for generating post-editing has been revised by GPT-4 iteratively five turns. We pick up a prompt that generates post-editing with the highest COMET-23 score in randomly picked 64 samples as the final prompt. The prompts are shown in Table A.1.

System:	As an expert in medical translation from English to Japanese, your task is to analyze the provided text pairs and identify discrepancies that affect the translation's accuracy and clarity. For each English source and Japanese translation pair, consider the following error types: 1. mis-translation: Note instances where the Japanese translation fails to capture the precise or contextual meaning of the English text. 2. syntax error: Identify any structural issues in the Japanese sentences that lead to grammatical errors or disrupt the natural flow of reading. 3. lexical choice: Point out where the choice of words in Japanese does not suit the medical context, is too general, or too ambiguous. 4. under-generation: Detect omissions in the Japanese text where essential information or terms from the English source are missing. 5. over-generation: Highlight any additional information or terms in the Japanese translation that are not present in the English source. For each identified error, provide a concise explanation and suggest a more accurate or appropriate translation. Focus on ensuring that the Japanese translation accurately reflects the medical terminology, context, and intent of the English source while maintaining a professional and reader-friendly tone. Utilize the reference for accurate terminology. Your goal is to enhance the translation to be as precise and clear as possible, making it a reliable medical resource.
User:	English source: Regular exercise can improve both of these qualities. Japanese translation: 通常の運動は、これらの性質を改善することができる。 Hints: 1. "通常の": mis-translation; 2. "both of": under-generation Reference: 定期的な運動によってその両方を向上させることができます。
Assistance:	Errors: 1. The word "通常の" in Japanese doesn't accurately convey the English meaning of "Regular". 2. The Japanese translation omits the specific emphasis on "both" of the qualities, making it less precise. Post-edited translation: 定期的な運動は、両方の性質を改善することができます。 (... more 7 demonstrations...)
User:	English source: $\{src\}$ Japanese translation: $\{mt\}$ Hints: {Spans of err} Reference: $\{ref\}$

Table 5 The prompt of gpt + mt + CoT + ref setting.