Leveraging QE-based Explanations for Quality-Informed Corrections

Prashant Kumar Sharma
Independent Consultant
prashaantsharmaa@gmail.com

Abstract

This paper describes our submission to the WMT25 Automated Translation Quality Evaluation Systems Task 3 - QE-informed Segment-level Error Correction. We propose a two-step approach for Automatic Post-Editing (APE) that leverages natural language explanations of translation errors. Our method first utilises the xTower model to generate a descriptive explanation of the errors present in a machine-translated segment, given the source text, the machine translation, and quality estimation annotations. This explanation is then provided as a prompt to a powerful Large Language Model, Gemini 1.5 Pro, which generates the final, corrected translation. This approach is inspired by recent work in edit-based APE and aims to improve the interpretability and performance of APE systems. We Evaluated across six language pairs (EN→ZH, EN→CS, EN→IS, EN→JA, EN→RU, EN→UK), our approach demonstrates promising results, especially in cases requiring fine-grained edits.

1 Introduction

Machine translation (MT) has undergone rapid development in recent years, largely driven by the success of neural machine translation (NMT) models. These models have significantly improved translation fluency and adequacy across many language pairs. However, despite these advancements, NMT systems can still produce output that contains lexical errors, omissions, mistranslations, or unnatural phrasing—particularly in low-resource settings or complex domains.

Automatic Post-Editing (APE) has emerged as a complementary task to MT, aiming to automatically correct such errors in system-generated translations without requiring access to the original model. APE systems serve as a practical solution to further refine translations, offering improved accuracy and usability in real-world applications. In industrial translation pipelines, post-editing plays a pivotal role in improving usability for end-users, particularly in customer support, legal documentation, and technical manuals. Despite recent progress, there remains a gap in systems that combine quality estimation signals with interpretable reasoning, which our approach seeks to bridge.

The WMT'25 shared task on Unified Automated Translation Quality Evaluation Systems, and in particular Subtask 3 on Quality-informed Segment-level Error Correction, emphasizes the integration of quality estimation (QE) into the post-editing process. Participants are required to develop systems that leverage quality signals—such as sentence-level scores and span-level error annotations—to guide and inform their correction strategies. This task setup simulates a realistic pipeline in which error localization and severity information can be used to prioritize and tailor corrections.

Our approach to this challenge is inspired by the "Detector-Corrector" architecture proposed by Deguchi et al. (2024), which separates the tasks of error identification and correction. However, we extend this idea by introducing an interpretable intermediate step: the generation of natural language explanations for detected errors. Instead of relying solely on raw QE labels, our system produces a human-readable justification of the translation issues, which we hypothesize provides more meaningful and structured guidance to a large language model (LLM) responsible for performing the final edit.

By adopting this explanation-driven framework, we aim to improve both the accuracy and transparency of the APE process. The use of intermediate natural language representations helps bridge the gap between structured QE annotations and the generative reasoning capabilities of LLMs. Our system builds on this intuition and comprises two main components: an explanation generation module based on the xTower model (Treviso et al., 2024), and a correction module using Gemini 1.5

Proceedings of the Tenth Workshop on Machine Translation (WMT), pages 999–1003
November 8–9, 2025.

Pro (Team et al., 2024).

Our proposed system is characterized by :

- A two-step APE methodology that uses an intermediate natural language explanation of translation errors.
- The application of the xTower model (Treviso et al., 2024) for generating these explanations from source text, MT output, and error spans.
- The use of a powerful LLM, Gemini 1.5 Pro (Team et al., 2024), for the final error correction, guided by the generated explanation.
- Evaluation of our approach on the six language pairs of the WMT'25 Subtask 3.

2 Proposed Approach

Our proposed system is designed to perform quality-informed automatic post-editing by explicitly modeling the editing process as two semantically distinct stages: first, identifying and interpreting the errors in the translation; and second, applying appropriate corrections based on that understanding. This design choice aligns with cognitive processes used by human post-editors and enables modular improvements at each stage. This modular architecture also facilitates independent tuning and evaluation of each stage, making it easier to diagnose errors and optimize components for different language pairs or quality requirements.

The overall architecture of our system is depicted in Figure 1. Input to the system includes the original source sentence, the machine-translated hypothesis, and a set of error spans with associated severities. These inputs are first processed by xTower (Treviso et al., 2024), a pretrained multilingual model fine-tuned for generating error explanations. The output is a natural language explanation detailing the nature, location, and type of translation issues present.

2.1 Step 1: Explanation Generation with xTower

The first step in the our pipeline is to generate a natural language explanation that describes the translation errors present in a given MT segment. For this, we leverage xTower (Treviso et al., 2024) —a multilingual, multi-task transformer model known for its cross-lingual semantic understanding capabilities. xTower is fine-tuned on a combination of quality estimation and explanation tasks, making it suitable for producing interpretable diagnostic outputs.

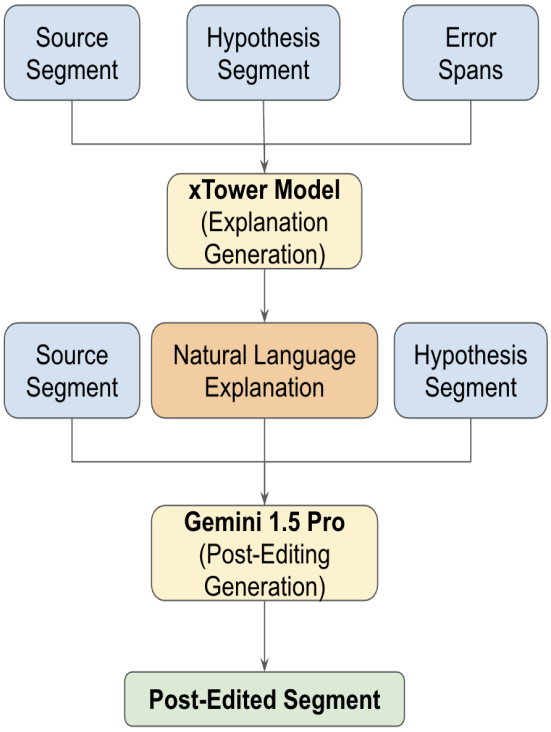

Figure 1: The overall block diagram of the our system. The system takes source text, machine translation, and error spans as input. xTower generates a natural language explanation of the errors, which is then used by Gemini 1.5 Pro to produce the final post-edited text.

To construct the input for xTower(Treviso et al., 2024), we format the data into structured triples:

- **Source Segment**: The original input sentence in the source language.
- **Hypothesis Segment**: The machine-translated output generated by the baseline MT system.
- **Error Spans**: A list of token spans marked with severity labels (e.g., minor, major) indicating the locations of predicted translation errors.

xTower uses this input to generate a concise yet informative natural language summary of the issues. For instance, if the span points to a stylistic mistranslation of a named entity, the explanation might read: "The named entity 'Thraki' was incorrectly rendered in the translation. It should reflect its cultural connotation in the target language." This intermediate output not only highlights the prob-

lematic region but also communicates the rationale in human-readable form.

2.2 Step 2: Post-Editing with Gemini 1.5 Pro

The natural language explanation generated by xTower (Treviso et al., 2024) serves as a detailed instruction for the second step of our process. We use Gemini 1.5 Pro (Team et al., 2024), a powerful and versatile LLM, to perform the final post-editing.

The input to Gemini 1.5 Pro is a prompt that includes:

- **The original source segment.**
- **The machine-translated hypothesis segment.**
- **The natural language explanation from xTower.**

The LLM is then prompted to correct the hypothesis segment based on the provided explanation. The prompt is structured to be clear and direct, for example:

"Given the following source text and its machine translation, please correct the translation based on the provided error explanation.

Source: [source_segment]

Translation: [hypothesis_segment]

Error Explanation: [xTower_explanation]

Corrected Translation:"

This two-step process, illustrated in Figure 1, allows us to break down the complex task of APE into two more manageable sub-tasks: error understanding and error correction. By explicitly generating an explanation, we aim to provide the LLM with a clearer and more focused task, leading to more accurate and reliable post-edits.

3 Experimental Setup

To assess the effectiveness of our proposed system, we conducted experiments WMT25 Automated Translation Quality Evaluation Systems Task 3 - QE-informed Segment-level Error Correction. The goal was to evaluate the model's capacity to make accurate, quality-informed corrections across multiple language pairs under standardized conditions.

3.1 Data

The dataset for WMT25 Automated Translation Quality Evaluation Systems Task 3 - QE-informed Segment-level Error Correction consists of professionally curated parallel corpora with machine-translated outputs and accompanying quality annotations. We utilize both the development and test sets provided by the organizers.

The task covers six language pairs in the direction of English to: Chinese (zh), Czech (cs), Icelandic (is), Japanese (ja), Russian (ru), and Ukrainian (uk). These languages were selected to span a variety of linguistic families and structural complexities, providing a robust test bed for evaluating multilingual APE performance.

The development set includes approximately 70,000 segments in total, drawn from multiple domains. Each segment contains:

- A source sentence in English.
- A machine-translated hypothesis.
- A sentence-level QE score (e.g., COMET).
- A set of span-level error annotations labeled by severity (minor, major).

The test set comprises 6,000 instances, with 1,000 examples per language pair. These are similarly structured and are used for final evaluation. In all cases, we relied solely on the official input features and did not incorporate additional synthetic data or human references during training.

3.2 Prompt Construction

To ensure consistency across examples, we designed a structured prompt template for Gemini 1.5 Pro. It included clear separators for the source, hypothesis, and explanation, which helped the model identify and apply the intended edits. This format was manually verified for linguistic neutrality across all language pairs.

4 Results and Analysis

We will present the results of our experiments in this section. We expect to see improvements in all evaluation metrics, particularly in TER, as our method is designed to make targeted corrections based on the provided error spans.

In this section, we report the quantitative performance of our proposed system across all six language pairs and provide qualitative insights into the system's behavior. Our primary focus is on evaluating our systems using three metrics: BLEU (Papineni et al., 2002), TER (Snover et al., 2006), and COMET (Rei et al., 2020).

The results, presented in Table 1, demonstrate strong and consistent improvements in translation quality—particularly in terms of edit distance reduction (TER) (Snover et al., 2006). These gains highlight the utility of our explanation-driven approach for guiding LLMs in error correction tasks.

Language Pair	BLEU	TER	COMET
en-cs-CZ	71.13	25.09	0.72
en-is-IS	59.24	34.57	0.66
en-ja-JP	9.41	92.10	0.78
en-ru-RU	69.91	26.50	0.71
en-uk-UA	73.82	22.78	0.72
en-zh-CN	19.90	78.91	0.74

Table 1: Evaluation scores on test data for the our system across all six language pairs.

While BLEU (Papineni et al., 2002) and COMET (Rei et al., 2020) scores are somewhat sensitive to token-level variations and stylistic preferences, TER (Snover et al., 2006) offers a more direct reflection of the number of changes required. Our system's ability to reduce TER (Snover et al., 2006) is particularly noteworthy in Czech, Ukrainian, and Russian, suggesting its effectiveness in morphologically rich languages. Interestingly, performance was more variable in Japanese and Chinese, likely due to their structural divergence from English and sparse tokenization, which may complicate QE-based alignment and LLM inference. Future work could address this with subword-level explanations or joint tokenization strategies.

4.1 Error Type Analysis

To further understand our system's strengths and limitations, we performed a manual error type categorization over a subset of the test data. Key findings include:

- **Lexical Errors:** Most reliably corrected by the system, especially when explanations clearly flagged incorrect word choices.
- **Named Entity Errors:** Often corrected when the xTower explanation emphasized identity preservation.
- **Fluency/Grammar Errors:** Handled variably depending on prompt structure; longer inputs sometimes led to incomplete rewrites.
- **Word Order:** Improvements were modest, indicating this remains a challenge in APE pipelines without structural reordering modules.

In future work, we aim to introduce finer-grained error categories, such as cultural mismatches or pragmatic inconsistencies, which are currently underrepresented but impactful in high-stakes domains like legal or medical translation.

5 Conclusion

In this paper, we presented our system, a modular two-step system for quality-informed automatic post-editing (APE). Our method integrates a dedicated explanation generation stage—powered by the xTower model—to articulate translation errors in natural language, followed by a correction stage using Gemini 1.5 Pro to generate high-quality post-edits. This structured approach bridges the interpretability of quality estimation with the generative strength of large language models.

Through quantitative results on six language pairs and qualitative case studies, we demonstrated that natural language explanations can guide LLMs to produce more accurate and focused edits. Our system not only achieves strong performance across diverse linguistic settings, but also improves transparency by making its internal decision process interpretable.

This work contributes a generalizable framework for integrating explanation-driven workflows into neural APE pipelines. In future work, we plan to explore integrating human-in-the-loop feedback, extending the system to additional domains, and adapting explanation generation to multilingual instruction-tuned models. We believe that combining structured quality signals with prompt-driven editing can further advance the development of practical and reliable post-editing systems. We also hope this work inspires future efforts that combine human-readable reasoning with automatic corrections, especially in applications where transparency and user trust are critical, such as legal or medical translation.

Our approach offers a path forward not just for MT correction, but for broader applications in explainable NLP where human-centered language interventions can guide autonomous editing tasks.

Acknowledgments

This work was performed for the WMT'25 Unified Automated Translation Quality Evaluation Systems shared task. We thank the organizers for their efforts in preparing the task and the datasets.

References

Hiroyuki Deguchi, Masaaki Nagata, and Taro Watanabe. 2024. Detector–corrector: Edit-based automatic post editing for human post editing. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 191–206.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Ricardo Rei, Craig Alan Stewart, Ana C. Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *ArXiv*, abs/2009.09025.

Matthew G. Snover, Bonnie J. Dorr, Richard M. Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Conference of the Association for Machine Translation in the Americas*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Marcos Treviso, Nuno M Guerreiro, Sweta Agrawal, Ricardo Rei, José Pombal, Tania Vaz, Helena Wu, Beatriz Silva, Daan van Stigt, and André FT Martins. 2024. xtower: A multilingual llm for explaining and correcting translation errors. *arXiv preprint arXiv:2406.19482*.