HW-TSC's submissions to the WMT 2025 Segment-level quality score prediction Task

Yuanchang Luo*, Jiaxin GUO*, Daimeng Wei*, Hengchao Shang, Zongyao Li, Zhiqiang Rao, Jinlong Yang, Zhanglin Wu, Xiaoyu Chen and Hao Yang
Huawei Translation Services Center, Beijing, China
{luoyuanchang1,guojiaxin1,weidaimeng}@huawei.com

Abstract

This paper presents the submissions of Huawei Translate Services Center (HW-TSC) to the WMT 2025 Segment-level quality score prediction Task. We participate in 16 language pairs. For the prediction of translation quality scores for long multi-sentence text units, we propose an automatic evaluation framework based on alignment algorithms. Our approach integrates sentence segmentation tools and dynamic programming to construct sentence-level alignments between source and translated texts, then adapts sentence-level evaluation models to document-level assessment via sliding-window aggregation. Our submissions achieved competitive results in the final evaluations of all language pairs we participated in.

1 Introduction

Recent advances in large language models (LLMs) (OpenAI, 2023; Touvron et al., 2023; Yang et al., 2024) have opened new possibilities for document-level machine translation (doc-mt) (Kim et al., 2019; Maruf et al., 2022; Fernandes et al., 2021). Leveraging their robust language generation capabilities and profound contextual understanding, LLMs can produce translations that are more natural, fluent, and semantically coherent. These models have demonstrated remarkable proficiency in processing long-form texts, thereby significantly enhancing the quality of document-level translation.

However, this approach also introduces several challenges. Since LLMs translate entire documents holistically rather than processing sentences sequentially, the output may suffer from issues such as over-translation (excessive paraphrasing) or under-translation (omissions). Furthermore, the absence of sentence-level alignment between source and target texts—combined with the inherent length of both—makes it difficult to assess translation quality accurately. Robust evaluation methods for document-level machine translation (MT) remain an unresolved critical problem.

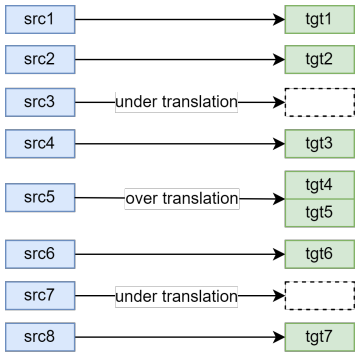

Figure 1: src_3 and src_7 lack corresponding translations in T, while src_5 aligns with a combined $tgt_4 + tgt_5$ segment.

While human evaluation remains the gold standard for assessing translation quality due to its nuanced understanding of language and context, it faces inherent limitations in scalability, subjectivity, and cost-efficiency, particularly for large-scale document-level translation tasks. Automated metrics like BERTScore (Zhang et al., 2019) and COMET (Rei et al., 2020a,b), though capable of capturing semantic nuances and demonstrating strong correlation with human judgments, are constrained by input length restrictions and their reliance on sentence-level alignment between source and reference texts. While (Vernikos et al., 2022) pioneered the adaptation of these metrics to document-level translation evaluation, its applicability remains severely constrained by its fundamental requirement for perfect sentence-level alignment among source texts, translations, and reference translations. This strict one-to-one correspondence prerequisite significantly limits its practical utility in real-world scenarios where such ideal alignments rarely exist. Recent attempts to leverage large language models (LLMs) as eval-

*These authors contributed equally to this work.

Proceedings of the Tenth Conference on Machine Translation, pages 969–973
November 8-9, 2025

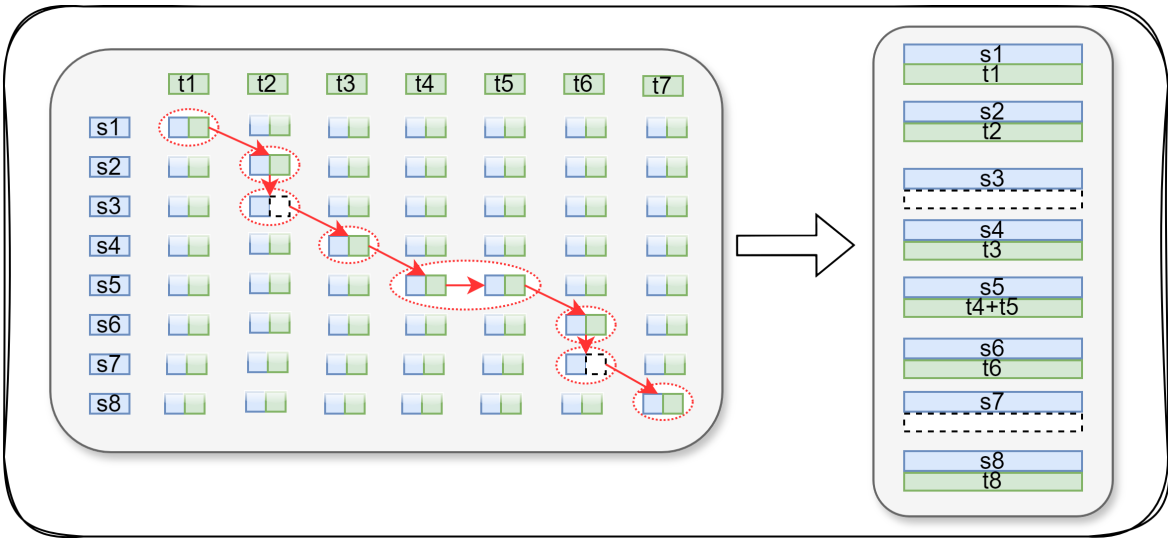

Figure 2: For the segmented text pair (8 source fragments and 7 target fragments), we first compute a full 8×7 score matrix using COMET KIWI to evaluate all possible pairwise alignments(subfigure a). We then apply dynamic programming to identify the optimal alignment path (visualized as the red trajectory in Figure). This optimization yields final sentence-level alignments, resulting in 8 properly aligned source-target pairs as demonstrated in subfigure (b).

uators through carefully designed prompts show promising alignment with professional human assessments across multiple dimensions including accuracy, fluency, and stylistic consistency (Gu et al., 2025). However, these methods suffer from high computational costs, sensitivity to training data biases, and instability across different prompts or model runs, raising concerns about their reliability and reproducibility for practical applications.

In this work, we employ an innovative alignment algorithm to automatically construct sentence-level alignment between source and translated texts. Our approach(Guo et al., 2025) involves: (1) sentence segmentation of source and target texts, (2) alignment metric computation, (3) anchoring of source text segmentation information, and (4) reconstructed target text segmentation (including merging and gap filling). By subsequently applying sliding-window-based sentence-level evaluation, we achieve document-level assessment effectiveness, thereby successfully adapting sentence-level pretrained model evaluation methods to document translation.

2 Approach

2.1 Alignment

Since our source text, translation, and reference translation are all document data, the sentence-level alignment between the source text and translation that we automatically construct can be divided into the following three parts:

- Sentence segmentation: Segment both original and translated texts into sentence sequences.
- Calculate alignment metrics: Measure alignment similarity between original and translated sentences using metrics like COMET KIWI (Rei et al., 2022) or LABSE (Feng et al., 2022).
- Reconstruct translated text segmentation: Based on the original text's segmentation, reconstruct the translated text's segmentation, involving possible merging or filling gaps. This is done using a dynamic programming algorithm.

As shown in Figure 2, for a source text S and its target translation T, we first perform sentence segmentation using spaCy [1], yielding m source sentences $S = (s_1, s_2, ..., s_m)$ and n target sentences $T = (t_1, t_2, ..., t_n)$. For these $m \times n$ sentence pairs, we compute a KIWI matrix $KIWI_{m \times n}$ using COMET KIWI. When $m = n$ with one-to-one correspondence, the diagonal path of this matrix should yield the maximum values. In document-level translation scenarios, the number of source segments and target segments typically differs

[1] https://spacy.io/

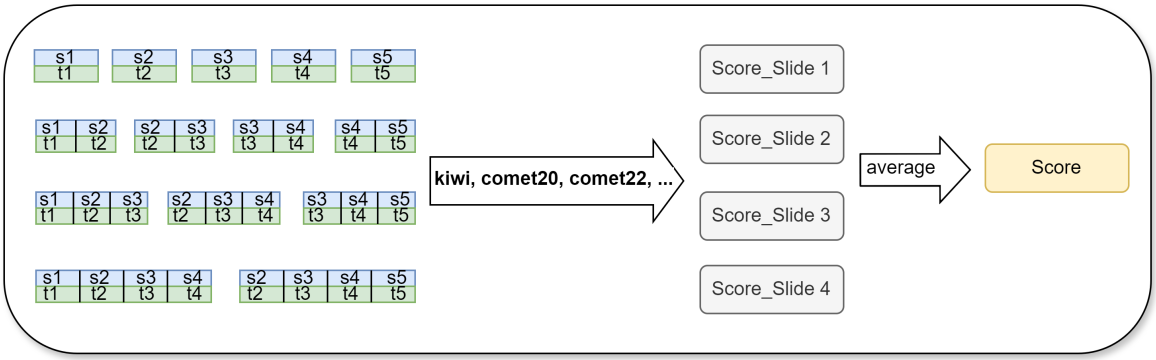

Figure 3: For the reconstructed source-target pairs, Compute Score Slide 1 on 5 original aligned pairs. Generate 4 concatenated pairs using window size 2 to calculate Score Slide 2 Generate 3 concatenated pairs using window size 3 to calculate Score Slide 3 Generate 2 concatenated pairs using window size 4 to calculate Score Slide 4. The final document-level metric is derived by averaging these four window-level scores, providing comprehensive coverage of local and contextual translation quality.

$(m \neq n)$. Nevertheless, we can identify an optimal alignment mapping $T = (t_1, t_2, ..., t_m) = F(s_1, s_2, ..., s_n)$ - represented as the optimal path in our framework - that maximizes the COMET KIWI score.

This alignment task can be abstracted as a path optimization problem: Given an $[mn]$ matrix where each cell (i, j) contains a score value, we seek the optimal path from $(0, 0)$ to $(m-1, n-1)$ under the following constraints:

- **Monotonicity Constraint**: y-coordinate must increase by exactly 1 at each step $(\forall t, y_{t+1} = y_t + 1)$. x-coordinate must increase by a non-negative integer $(\forall t, x_{t+1} \geq x_t)$
- **Boundary Conditions**: Path originates at the top-left corner $(0, 0)$ and terminates at the bottom-right corner $(m-1, n-1)$
- **Optimization Objective**: Maximize the cumulative score:

$$argmax_p \sum_{(x,y)\in p} matrix[x][y] \quad (1)$$

Using the dynamic programming algorithm, we can obtain a translation whose segmentation aligns one-to-one with the source text, as well as the segmentation information of the reference translation.

2.2 Sliding Evaluation

After obtaining the alignment information in the previous step, we follow a procedure similar to (Raunak et al., 2024), calculating sentence-level scores using a sliding window approach. As illustrated in Figure 3, for m source sentences $S = (s_1, s_2, ..., s_m)$ and their aligned translations $T' = (t'_1, t'_2, ..., t'_m)$, given a window size n, we compute m groups of sentence-level evaluation metrics, each incorporating $n-1$ preceding sentences as contextual information. The mean of these scores serves as the document-level evaluation result, expressed formally as follows:

$$\frac{1}{n}\sum_{i=1}^{n} f_i(S, T') \quad (2)$$

Where f_i corresponds to the Slide Score measured when the window is i, corresponding to Score Slide i in Figure 3.

3 Results

We participated in all language pair competitions within the Segment-level Quality Score Prediction Task, which included a total of 16 language pairs. After the alignment phase, we obtained new sentence-level text pairs corresponding to each paragraph text pair. At this point, we conducted respective predictions using wmt22-cometkiwi-da(**ASD-KIWI**), wmt23-cometkiwi-da-xl(**ASD-KIWI-XL**), and wmt23-cometkiwi-da-xxl(**ASD-KIWI-XXL**), with the results shown in Table 1. As shown in the Table 1, **ASD-KIWI-XL** demonstrates superior correlation to **ASD-KIWI** across most of the 16 language pairs, indicating that post-alignment sentence-pair quality scoring plays a critical role. While larger parameter models generally achieve better performance (as evidenced by KIWI-XL's gains), this trend is not absolute—**ASD-KIWI-XXL** fails to further outperform **ASD-KIWI-XL**.

Languages Pairs	ASD-KIWI	ASD-KIWI-XL	ASD-KIWI-XXL	ASD-KIWI-ENSEMBLE
EN-ZH	0.6800	0.6467	0.5600	**0.7467**
CS-UK	0.7382	0.5418	0.7164	**0.7818**
EN-KO	0.7067	0.7600	0.7267	**0.7733**
EN-IT	0.7169	0.600	0.5077	**0.7169**
EN-ET	0.7018	0.7164	0.6436	**0.7455**
EN-BHO	0.9316	0.9031	0.8348	**0.9316**
EN-IS	**0.8667**	0.7667	0.7400	0.7933
EN-SR	0.8974	0.8575	0.8519	**0.9031**
CS-DE	0.719	0.5820	0.6676	**0.7418**
EN-RU	0.6710	0.6017	0.3853	**0.8355**
EN-JA	**0.7933**	0.6533	0.4933	0.7667
EN-AR	0.8551	0.8696	0.7681	**0.8551**
EN-UK	0.7524	0.7238	0.5048	**0.8190**
EN-MAS	**0.7628**	0.5652	0.5889	0.5968
EN-CS	0.5942	0.6087	0.5362	**0.6957**
JA-ZH	**0.7245**	0.5042	0.6177	0.6978

Table 1: Results for 16 Languages Pairs in the Segment-Level Quality Score Prediction Task

To leverage both models, we propose an ensemble method that averages the per-sentence scores of **ASD-KIWI** and **ASD-KIWI-XL**. Empirical results confirm that **ASD-KIWI-ENSEMBLE** achieves the best overall performance.

4 Conclusion

This paper presents the methodology behind HW-TSC's submission to the WMT 2025 Segment-Level Quality Score Prediction Task. Our approach integrates sentence segmentation tools and dynamic programming algorithms to construct sentence-level alignments between source and translated texts, then adapts sentence-level evaluation models to document-level assessment through sliding-window aggregation. By incorporating an ensemble strategy, our method achieved the highest correlation scores across all 16 languages in this task.

References

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 878–891. Association for Computational Linguistics.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. 2021. Measuring and increasing context usage in context-aware 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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6467–6478. Association for Computational Linguistics.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Jiaxin Guo, Daimeng Wei, Yuanchang Luo, Xiaoyu Chen, Zhanglin Wu, Huan Yang, Hengchao Shang, Zongyao Li, Zhiqiang Rao, Jinlong Yang, and 1 others. 2025. Align-then-slide: A complete evaluation framework for ultra-long document-level machine translation. *arXiv preprint arXiv:2509.03809*.

Yunsu Kim, Duc Thanh Tran, and Hermann Ney. 2019. When and why is document-level context useful in neural machine translation? In *Proceedings of the Fourth Workshop on Discourse in Machine Translation, DiscoMT@EMNLP 2019, Hong Kong, China, November 3, 2019*, pages 24–34. Association for Computational Linguistics.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. 2022. A survey on document-level neural machine translation: Methods and evaluation. *ACM Comput. Surv.*, 54(2):45:1–45:36.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Vikas Raunak, Tom Kocmi, and Matt Post. 2024. SLIDE: reference-free evaluation for machine trans-

lation using a sliding document window. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Short Papers, NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 205–211. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020a. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2685–2702. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020b. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation, WMT@EMNLP 2020, Online, November 19-20, 2020*, pages 911–920. Association for Computational Linguistics.

Ricardo Rei, Marcos V. Treviso, Nuno Miguel Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte M. Alves, Luísa Coheur, Alon Lavie, and André F. T. Martins. 2022. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 634–645. 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, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 118–128. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *CoRR*, abs/2407.10671.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.