Align-then-Slide: A complete evaluation framework for Ultra-Long Document-Level Machine Translation

Anonymous ACL submission

Abstract

In recent years, large language models (LLMs) have significantly advanced document-level translation quality, leveraging their powerful text-generation and context-understanding capabilities. However, since document-level translation generates outputs holistically rather than sentence-by-sentence, it often suffers from over-translation, under-translation, and the lack of sentence-level alignment information, posing substantial challenges for quality assessment. Existing evaluation methods (e.g., BERTScore, COMET) struggle with long-input constraints, making them impractical for direct application to document-level translation. To address these issues, 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. Experiments show that our method efficiently and accurately evaluates document-level translation quality, offering a reliable tool for future research.

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.

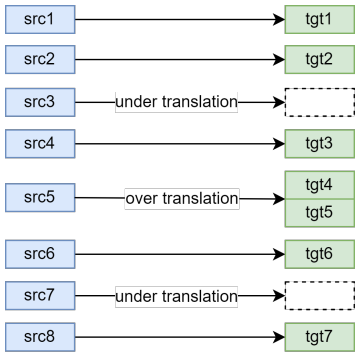

Figure 1: src_3 and src_7 lack corresponding translations in T, while src_5 aligns with a combined $tgt_4 + tgt_5$ segment.

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., 2020), 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

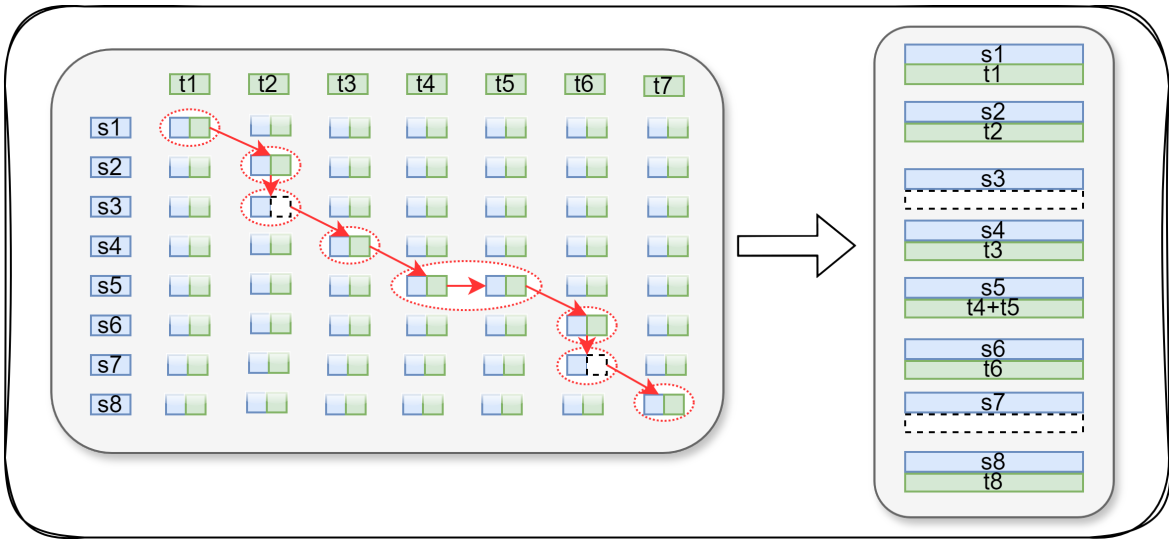

Figure 2: For the segmented text pair (8 source fragments and 7 target fragments), we first compute a full 8×7 score matrix using COMET KIWI to evaluate all possible pairwise alignments(subfigure a). We then apply dynamic programming to identify the optimal alignment path (visualized as the red trajectory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

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 evaluators 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 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., 2020) 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

[1] https://spacy.io/

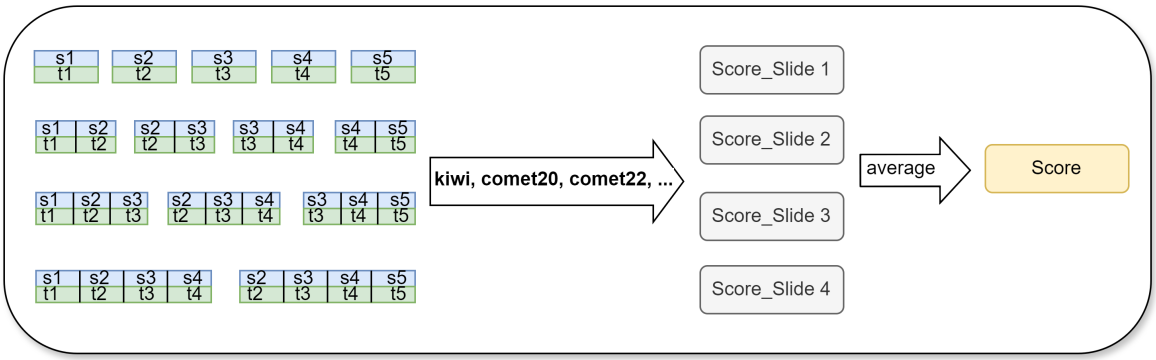

Figure 3: For the reconstructed source-target pairs, Compute Score Slide 1 on 5 original aligned pairs. Generate 4 concatenated pairs using window size 2 to calculate Score Slide 2 Generate 3 concatenated pairs using window size 3 to calculate Score Slide 3 Generate 2 concatenated pairs using window 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.

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 ($m \neq n$). 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:

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

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 Paper A, 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:

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

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

3 Experiments

We conducted experiments on the test set from the IWSLT2017 translation task 6[2], comprising parallel documents from TED talks. Our experiments encompassed eight language pairs: English-German, English-French, English-Japanese, and English-Chinese in both directions.

we employed a suite of Qwen models[3] (ranging from 7B to 72B parameters) for translation generation Given our direct utilization of these large language models for document-level translation, we set the maximum number of new tokens to 16,384

[2]https://wit3.fbk.eu/2017-01-d/
[3]https://huggingface.co/Qwen

	model	ASD-20	ASD-22	ASD-KIWI
xx2en	Qwen2.5-7B-Instruct	0.4939	0.8293	0.8184
	Qwen2.5-14B-Instruct	0.4906	0.8304	0.8187
	Qwen2.5-32B-Instruct	0.5041	0.8345	0.8192
	Qwen2.5-72B-Instruct	0.5181	0.8385	0.8207
en2xx	Qwen2.5-7B-Instruct	0.0207	0.691	0.7325
	Qwen2.5-14B-Instruc	0.3361	0.7018	0.8138
	Qwen2.5-32B-Instruct	0.3502	0.7125	0.8189
	Qwen2.5-72B-Instruct	0.3746	0.7255	0.8287

Table 1: Results for different model sizes in the test set

to accommodate lengthy document inputs while maintaining computational feasibility.

In our evaluation framework, we compute COMET scores for each aligned sliding window segment and average them to derive document-level metrics - specifically ASD-20, ASD-22, and ASD-KIWI. These systematically designed metrics enable rigorous validation of our alignment-based assessment approach.

4 Results

Table 1 reveals that for the $xx2en$ translation direction, the Qwen model series exhibit relatively minor variations across all three $ASD-20$ metrics regardless of parameter size. However, a consistent (though modest) positive correlation between model scale and metric scores can be observed. For instance, the Qwen2.5-72B-Instruct model achieves an $ASD-20$ score of 0.5181, outperforming its 7B counterpart (0.4939) by a margin of 0.0242 - demonstrating the expected scaling trend despite the generally small performance gaps.

In the $en2xx$ translation direction, we observe significantly larger performance gaps among Qwen models of different scales. The Qwen2.5-7B-Instruct model achieves only a 0.0207 ASD-20 score, with manual inspection revealing frequent hallucinations and severe under-translation in its outputs. This deficiency progressively diminishes with increased model size: the 14B variant shows a substantial 0.3154 point improvement (0.3361 vs 0.0207), while the gap between 14B and 72B models narrows to just 0.0385 (0.3361 vs 0.3746). These results suggest that for $en2xx$ document translation, Qwen models require at least 14B parameters to produce adequate quality - a finding consistent with practical deployment experience.

In summary, our proposed evaluation method consistently captures the expected scaling law - larger models achieve better performance - across both translation directions. The particularly pronounced quality gap in $en2xx$ translation provides strong empirical validation of our framework's sensitivity to model capability differences in document-level translation assessment.

5 Conclusion

In this paper, We propose a novel solution combining sentence segmentation tools and dynamic programming algorithms to address the sentence-level misalignment problem among source texts, translations, and reference translations in document-level translation. Enable effective migration of document-level translation evaluation to sentence-level assessment through our aligned sentence pairs, with our sliding-window-based approach being particularly suitable for document translation evaluation. In a multi-language text test set, the effectiveness of our method is verified against large language models with different parameter sizes.

Limitations

While our proposed method successfully bridges sentence-level evaluation metrics to document-level translation assessment, its performance critically depends on the initial alignment step. This process requires computing an m×n score matrix, where m and n represent the numbers of source and target segments respectively. As document length increases, the computational resources needed for this matrix grow quadratically ($O(m \times n)$). We identify this scalability challenge as a key limitation to be addressed in future work.

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.

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.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. 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.

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.