Context Repair with Large Language Models for Document-level Neural Machine Translation

Anonymous ACL submission

Abstract

Current neural machine translation models generate translation output sentence-by-sentence, where each translation procedure is carried out independently. This sentence-level decoding strategy results in an inherent issue of incoherence. Consequently, considerable effort has been dedicated to document-level machine translation to mitigate the incoherence problem. In this work, we propose a simple and effective technique to repair document context by leveraging the power of large language models. The document-level translation task is decomposed into a sentence-level translation task and a contextual information repair task. we first employ a conventional sentence-level translation model to generate sentence-level translation outputs. Then, we pair these outputs with their corresponding translation references to create few-shot examples. Finally, we utilize a large language model along with these few-shot examples to perform context repair for the test sentences. Experimental results on the Bilingual Web Books test set demonstrate the effectiveness of the proposed approach in document-context translation. Besides, the approach also works with other methods. Further analysis and human evaluation results indicate that the proposed approach outperforms the baseline model in terms of human preference.

1 Introduction

In recent years, conventional sentence-level machine translation has achieved remarkable progress. However, it often fails to capture the intricate relationships and dependencies between sentences, leading to inconsistencies and incoherent errors in the translation output. As a result, document-level machine translation, which aims to translate an entire document instead of individual sentences from one language to another, has received much attention over the years (Hardmeier et al., 2012; Wang et al., 2017; Jean et al., 2017; Sun et al., 2022). Despite great promise, generating long and coherent sequences in document-level machine translation remains a challenge.

In recent years, large language models (LLMs) have shown a remarkable ability across a wide range of natural language processing (NLP) tasks (Radford et al., 2019; Ouyang et al., 2022). LLMs exhibit superior translation capabilities for sentence-level (Jiao et al., 2023; He et al., 2023) and document-level translations (Hendy et al., 2023; Wang et al., 2023).

To further explore the cross-linguistic capability of the LLMs, Raunak et al. (2023) and Chen et al. (2023) propose employing the zero-shot prompting strategy to refine the sentence-level translation output. Koneru et al. (2023) proposes to adapt LLMs as automatic post-editors. Wu et al. (2024) explores fine-tuning large language models on the document-level parallel corpus and uses the fine-tuned model to assist document-level translation and post-translation editing tasks

In this work, we propose a simple and effective technique that uses LLMs with few-shot prompt learning to repair the document context for document-level machine translation. The main idea is to exploit the powerful capabilities of LLMs in capturing long-range dependencies and understanding complex linguistic structures across sentences.

Specifically, we first use a conventional sentence-level translation model to generate sentence-level translation outputs. Then, we pair these outputs with their corresponding translation references to create few-shot examples {document translation, document reference}. Finally, we utilize a large language model along with these few-shot examples to perform context repair for the test sentences. The LLM addresses the context of sentence-level translation in the test set, thereby generating improved document-level translations.

Experimental results on the Bilingual Web Books (BWB) test set demonstrate the effectiveness of the proposed approach in document-context

translation. The proposed approach can effectively identify and rectify inconsistencies and incoherence errors in the translation output to improve overall translation quality. We also conduct extensive analyses and human evaluations to confirm that the proposed approach outperforms the baseline model in terms of human preference.

The advantages of our approach are fourfold: (1) It is versatile and can be applied to both conventional system-level and document-level machine translation systems. (2) By employing few-shot learning, our approach does not require a large amount of costly document-level parallel data. (3) The two-pass decoding in the proposed approach does not require much extra time overhead. (4) Our method can be applied to other methods to achieve better results.

2 Background

2.1 Machine Translation

Machine translation is a subfield of computational linguistics and artificial intelligence that focuses on developing algorithms and models capable of automatically translating text or speech from one language to another. Over the years, various approaches have been explored, including rule-based, statistical, and, more recently, neural machine translation techniques. Generally, given a translation model M trained on the bilingual sentence pairs, the translation process for a document $D = [s_1, s_2, ..., s_n]$ comprising a set of n sentences is divided into individual sentence translations. The sentence translation producer can be formulated as:

$$t = P(s|\theta_M) : s \in D^s \quad (1)$$

where θ_M represents the parameter of the model M, and t is the target sentence translation. Finally, we get the document translation $D^t = [t_1, t_2, ..., t_n]$, consisted of n target sentences.

2.2 Large Language Model

Few-shot prompting(Brown et al., 2020) has emerged as a promising paradigm in large language models, enabling these models to adapt to new tasks and domains with limited labeled data. In contrast to traditional supervised learning approaches that require extensively annotated datasets, few-shot prompting capitalizes on the pre-training and fine-tuning strategies employed by large language models. Few-shot prompting ability to generalize from limited data has significant implications for various natural language processing tasks, particularly in low-resource settings or specialized domains where labeled data is scarce. Concretely, few-shot prompting first converts each test data to a prompt and then generates the response by feeding the prompt to the pre-trained LLM.

3 Context Repair with LLMs

In this section, we introduce the proposed approach to context repair in document-level machine translation, leveraging the capabilities of LLMs. Our approach consists of three main steps:

Document translation. We translate the source documents in a sentence-by-sentence manner, utilizing a pre-trained translation model to generate document translation as described in Section 2.1.

Few-shot prompting preparation. We pair the document translation output with the corresponding reference to construct the few-shot prompting template. This template serves as a means for LLMs to understand the desired context repair task.

Context Repair. For the test source document, we first obtain the initial document translation and then convert the initial translation to a prompt and feed the prompt to the pre-trained LLM.

Using the few-shot examples, the LLM learn to identify and rectify inconsistent and incoherent errors in the initial translation, ensuring that the final document translation is both contextually accurate and coherent.

4 Experiment

4.1 Setup

Dataset. We conduct experiments on large-scale Chinese-to-English document-level parallel corpus BWB (Jiang et al., 2022). BWB consists of Chinese web novels across multiple genres (sci-fi, romance, action, fantasy, comedy, inter alia), and their corresponding English translations crawled from the Internet.

Model. We conduct experiments on both closed-source and open-source LLMs.

- VICUNA (Chiang et al.) (vicuna-13b-v1.5-16k[1]) An open source available strong Chatbot, fine-tuned from Llama 2 with supervised instruction fine-tuning and has a long contextual input window.

[1] https://huggingface.co/lmsys/vicuna-13b-v1.5-16k

System	few-shot	BLONDE	ENTITY	TENSE	PRON	DM	Avg.
MT-D (Jiang et al., 2022)	–	34.37	43.51	79.28	79.35	67.08	67.31
VICUNA							
BASELINE	–	24.66	14.24	75.25	75.96	66.48	57.98
+ Context repair w/o. source	3	22.12	15.47	69.21	69.69	57.31	52.92
	5	22.23	14.00	69.88	68.88	59.25	53.00
+ Context repair w. source	3	24.44	17.95	72.13	73.24	61.17	56.12
	5	26.53	18.54	77.04	77.41	65.05	59.51
CHATGPT							
BASELINE	–	34.74	43.53	79.08	79.62	**69.04**	67.82
+ Context repair w/o. source	3	33.40	**46.56**	79.87	81.73	66.84	68.75
	5	33.61	46.27	80.20	81.76	66.68	68.73
+ Context repair w. source	3	35.39	43.96	80.78	82.26	68.78	68.95
	5	35.62	45.00	80.81	82.67	68.61	69.27
GPT-4							
BASELINE	–	34.38	41.41	77.63	78.80	68.53	66.59
+ Context repair w. source	3	35.10	43.84	80.52	83.74	67.80	68.98
	5	35.65	45.98	81.28	84.22	68.33	69.95
ITERATIVE REFINEMENT							
BASELINE	–	32.55	41.64	76.63	79.99	67.87	66.53
+ Context repair w. source	3	37.22	44.51	81.08	84.86	68.38	69.71
	5	**38.27**	46.55	**81.58**	**84.99**	68.75	**70.47**

Table 1: Document-level evaluation on the BWB test set in terms of BLONDE, entity (ENTITY), tense (TENSE), pronoun (PRON), discourse marker (DM). "Avg." denotes the average score of ENTITY, TENSE, PRON, and DM.

- CHATGPT (text-davinci-003[2]) A strong but closed-source LLM developed by OpenAI. This model serves as a strong baseline, achieving or even surpassing the best submission of WMT22 in many directions (He et al., 2023).
- GPT-4 (gpt-4-0125-preview[3]) A stronger than ChatGPT but closed-source LLM developed by OpenAI. It has stronger instruction-following ability and stronger context-understanding ability than ChatGPT.

Comparison Systems. We also compare our approach with the following systems: 1) MT-D (Jiang et al., 2022), which splits the source document into slices and then translates slice by slice. 2) ITERATIVE REFINEMENT (Chen et al., 2023), which introduces iterative refinement for the translation output of the system.

Prompt Format. Due to the limitation of prompt length, we simplified the prompt sample by using continuous n sentences to construct a pseudo-document. We follow Zhang et al. (Zhang et al., 2023) to construct the few-show prompt "[input] t_1 [output] r_1 ... [input] t_n [output] r_n [input] t [output]" where the $t_1...t_n$ are pseudo-document translation, the $r_1...r_n$ are the pseudo-document reference, and t is the pseudo-document translation to be revised. We name the approach "context repair without source".

We also validate the proposed approach through the few-shot prompting strategy with the source language sentence guidance. The format of the few-show prompt is "[src] s_1 [input] t_1 [output] r_1 ...[src] s_n [input] t_n [output] r_n [src] s [input] t [output]" where the $s_1...s_n$ are the source sentences of the pseudo-document. We name the approach "context repair with source".

Metrics and Evaluation. The evaluation metrics used in our experiments are as follows:

- BLONDE (Jiang et al., 2022): An automatic Evaluation Metric based word match for Document-level Machine Translation.
- ENTITY (Jiang et al., 2022): Compute entity identity by FastText precies, recall and F1 socre between reference and target sentences.

[2] https://openai.com/chatgpt
[3] https://openai.com/gpt-4

- TENSE (Jiang et al., 2022): Identify seven types of tense spans in the text by models, and then calculate the corresponding F1 score. Indicating the consistent tense in the translation.

- PRON (Jiang et al., 2022): The feature pronoun, computes three predefined groups of pronouns F1 score in translation.

- DM (Jiang et al., 2022): The feature discourse maker computes four predefined groups of discourse marker types F1 score.

4.2 Main Results

Table 1 reports the experimental results in terms of discourse-related metrics. We can find that

- Using more examples (5-shot vs. 3-shot) in the few-shot prompting strategy enhances the discourse performance of the proposed method.

- Incorporating source sentences in the few-shot prompting strategy enables the performance of the proposed approach to surpass the baseline model in all settings.

- Cooperating with ITERATIVE REFINEMENT, our approach achieves the best performance on average, indicating that our approach also works with other methods.

- The improvement of our approach with the GPT-4 setting (69.95-66.59=3.36) is higher than that under ChatGPT (69.27-67.82=1.45) and Vicuna (59.51-57.98=1.53) models, which indicates that the GPT-4 model has a stronger ability to document context information for the machine translation task.

4.3 Analysis

We observe the discourse maker computing score drop in our approach. We attribute it to the calculation of DM metric. For example, in one case, the reference text incorporates three discourse marker words: “when," “then," and “after". The ChatGPT prediction only generates “when", and our approach produces two rephrasings: “when" and “but."

As our approach contains an additional occurrence of “but" compared to the ChatGPT, and there is no discourse marker word (Jiang et al., 2022) of the same type in the reference, the DM precision score of our approach is lower than that of the ChatGPT (For details on the discourse marker words and the calculation of DM score, see (Jiang et al., 2022)). However, it is worth noting that the occurrence of “but" in the rewrite should correspond to “instead" in the reference during translation. However, the BLONDE metric does not consider it as a discourse marker word. That is the reason why our approach has a lower DM score than the baseline model.

We conduct a human evaluation to compare the proposed approach (utilizing a 5-shot prompting strategy with source sentence, and LLM is ChatGPT) to the corresponding baseline model. Given the source Chinese sentence, the annotators were asked to determine the higher quality translation among two translations, i.e., the translation outputs generated by the proposed approach and the baseline model. The generation system of the translation is agnostic to the annotator, and the order of the two translations is random.

Figure 1 shows the results of the human evaluation study. We find that the proposed approach achieves better translations at 52.43% compared to the baseline model, confirming the effectiveness of the proposed approach.

Figure 1: A human evaluation was conducted on the BWB test set, comparing the proposed approach (utilizing a 5-shot prompting strategy with source sentence) to the corresponding baseline model.

5 Conclusion

In this study, we propose to address the discrepancies and inherent inaccuracies stemming from sentence-level machine translation by employing LLMs to repair translation outputs. Our approach involves conducting experiments on the BWB dataset and evaluating the results using both automated metrics, such as BLONDE, and human assessments. The experimental results demonstrate the effectiveness of our proposed method in enhancing discourse phenomena in document translation.

6 Limitations

Our context repair approach utilizing LLMs for document-level translation presents the following limitations:

- **Context window expansion**: The context window constrains the quantity of examples in few-shot prompting. Recently, a variety of methods aimed at expanding the context window have been developed, such as NTK-Aware Scaled RoPE[4]. And the impact of these methods on document-level translation needs to be explored.

- **More advanced metrics**: The metrics employed in this study are rule-based. Future research should explore the use of semantic-based metrics for document-level translation.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2023. Iterative translation refinement with large language models. *arXiv preprint arXiv:2306.03856*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Christian Hardmeier, Joakim Nivre, and Jörg Tiedemann. 2012. Document-wide decoding for phrase-based statistical machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023. Exploring human-like translation strategy with large language models. *arXiv preprint arXiv:2305.04118*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Sebastien Jean, Stanislas Lauly, Orhan Firat, and Kyunghyun Cho. 2017. Does neural machine translation benefit from larger context? *arXiv preprint arXiv:1704.05135*.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*.

Sai Koneru, Miriam Exel, Matthias Huck, and Jan Niehues. 2023. Contextual refinement of translations: Large language models for sentence and document-level post-editing. *arXiv*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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

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*.

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*.

Longyue Wang, Zhaopeng Tu, Andy Way, and Qun Liu. 2017. Exploiting cross-sentence context for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831.

[4] https://www.reddit.com/r/LocalLLaMA/comments/14lz7j5/ntkaware_scaled_rope_allows_llama_models_to_have/

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

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. *Proceedings of the 40th International Conference on Machine Learning*.