DUTNLP System for WMT23 Discourse-Level Literary Translation

Anqi Zhao[1], **Kaiyu Huang**[2], **Hao Yu**[1], **Degen Huang**[1*]
[1]Dalian University of Technology, Liaoning, China
[2]Institute for AI Industry Research (AIR), Tsinghua University, Beijing, China
user_zaq@mail.dlut.edu.cn; huangdg@dlut.edu.cn;

Abstract

This paper details the submission from the DUTNLP Lab for the WMT23 Discourse-Level Literary Translation in Chinese to English translation direction under unconstrained conditions. Our primary system aims to harness a large language model with various prompt strategies, allowing for a comprehensive exploration of the potential capabilities of large language models in discourse-level neural machine translation. Moreover, we apply detailed data preprocessing methods to filter bilingual data, which proves to be beneficial. Additionally, we assess a widely used discourse-level machine translation model, G-transformer, using different training strategies. In our experimental results, the method employing large language models achieves a BLEU score of 28.16, whereas the fine-tuned method scores 25.26. These findings indicate that selecting appropriate prompt strategies based on large language models can significantly enhance translation performance compared to traditional model training methods.

1 Introduction

The DUTNLP Lab is actively participating in WMT23 Discourse-Level Literary Translation, focusing on Chinese to English translation direction. As observed, prompting large language models (LLMs) has led to outstanding performance across a range of natural language processing (NLP) tasks (Chowdhery et al., 2022; Goyal et al., 2023; Chung et al., 2022). So our research involves experimenting with various prompts and in-context learning strategies, utilizing large language models. Additionally, we conduct experiments to explore the impact of sentence length and data preprocessing methods on translation results.

Our research is primarily anchored in the gpt-3.5-turbo model (Brown et al., 2020), renowned for its outstanding language generation capabilities spanning various domains, from writing to conversations. This model excels at producing natural and fluent text with simple prompts, making it accessible even to individuals without extensive technical knowledge.

Intriguingly, for crafting effective prompts to stimulate the machine translation capability of the large model, we take inspiration from gpt-3.5-turbo. We actively interact with it to derive prompts that can boost translation performance, resulting in the identification of three candidate translation prompt templates. Our evaluation of these prompts in the discourse-level translation task indicates their overall effectiveness, with minor performance variations.

Recognizing the substantial impact of data quality on translation performance, we employ cleaner development corpora for our main experiments. When utilizing large pre-trained models, we conduct a data filtering process through off-the-shelf tools and manual rule-based approaches. Further details will be seen in Session 2.

Given the inherent randomness and flexibility in translations generated by large models, aligning the output with the source text can be challenging. To tackle this challenge, we develop scripts to identify segments with alignment errors and subsequently apply manual corrections for rectification.

To sum up, our contributions can be outlined as follows:

- We have carefully crafted a prompt that has led to a notable performance of 28.16 BLEU (Papineni et al., 2002) on our dataset. This accomplishment suggests a significant improvement over standard document-level machine translation models, including the G-transformer model (Bao et al., 2021), trained with various strategies.
- We have conducted a series of meticulously controlled experiments to systematically in-

*Corresponding authors

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 296–301
December 6–7, 2023.

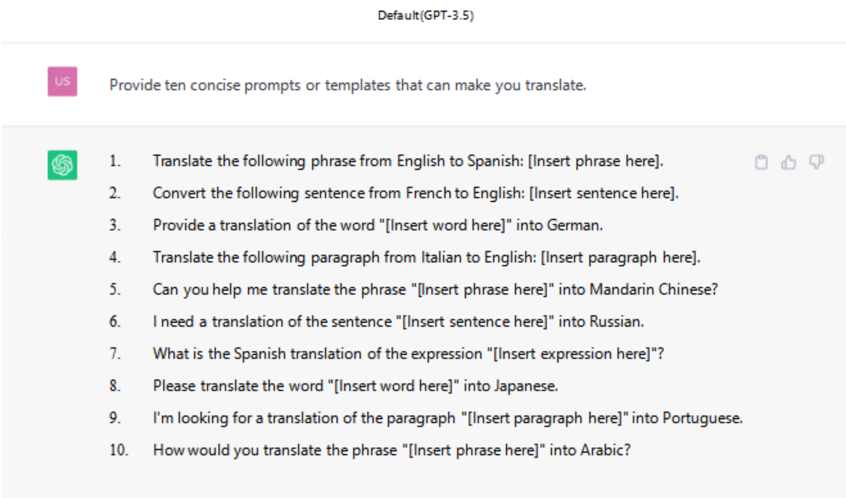

Figure 1: Prompts advised by gpt-3.5-turbo for machine translation.

vestigate the impact of different prompt strategies, batch sentence quantities, and tokenizer methods on the performance of the gpt-3.5-turbo model when apply to the Chinese-to-English discourse-level translation task.

This paper is structured as follows: Section 2 describes the data pre-processing strategies, followed by the details of our method in Section 3. Section 4 presents the experimental results and analysis, and we draw conclusions in Section 5.

2 Data Processing

Contrary to the conventional fine-tuning approach on large language models, our method utilizes a large pre-trained language model combined with prompts. In other words, Our primary experiment do not require further model training. Therefore, we conduct experiments using only a small portion of the development dataset.

Since the data quality significantly impacts our final translation performance, we adopt both traditional data processing methods and manual rules for filtering. The pre-processing strategies are as follows:

- Extract the discourse-level data from the text data with HTML tags and filter out duplicated sentence pairs.
- Filter out sentences containing illegal and invisible characters, like certain emoji symbols, as they may cause alignment issues.
- Normalize punctuation using Moses scripts (Koehn et al., 2007) for English and Chinese. Chinese text is separately segmented by Jieba tool.
- For Chinese, convert full-width format to half-width format and traditional Chinese characters to simplified ones.

	Translation Prompt
TP1	Translate the following sentences "[Insert text here]" from [SRC] to [TGT].
TP2	These sentences "[Insert text here]" are in [SRC] and can be translated to [TGT] as follows:
TP3	Please provide translations of these sentences "[Insert text here]" into [TGT].

Table 1: Candidate translation prompt.

3 Method

To unlock the full potential of large language models, we introduce an innovative approach by seeking guidance from gpt-3.5-turbo for the creation of effective machine translation prompts (Jiao et al., 2023). Specifically, we pose the following query: 'Provide ten concise prompts or templates that can prompt translation.'

The obtained results are shown in Figure 1. Upon observation, we note that the generated prompts are reasonable and similar. Consequently, we consolidate them into three sets of candidate templates, as illustrated in Table 1, where [SRC] and [TGT] represent the source and target language of translation.

In previous studies concerning discourse-level machine translation, it is evident that factors such as varying discourse lengths (Wang and Cho, 2019; Raffel et al., 2019) and different segmentation granularities (Koehn, 2005; Sennrich et al., 2016) can significantly impact translation performance. Consequently, we design a series of comparative experiments to investigate these aspects. Specifically, we segment the document texts into sizes of k and analyze the effects of different text lengths on machine translation performance in our experimental results.

During the segmentation of document text, our goal is to achieve an equitable distribution of text segments and prevent a situation where only a few isolated sentences remain at the end of a document. To address this, we devise a text segmentation algorithm that preserves the data while also ensuring that the number of text portions between segments is as uniformly distributed as possible. The aim is to minimize variance in sentence counts, as illustrated below.

The main strategy is as follows: for a document containing n lines of text, it undergoes slicing based on a specified size of m lines, where the quotient is denoted as p and the remainder as q. If there is a remainder ($q \neq 0$), it indicates the need to slice the text into $n/p + 1$ segments. This results in a new quotient, k, and a new remainder, t. Consequently, the last t segments are allocated a line count of $k + 1$, while the rest of the segments maintain a line count of k.

In traditional machine translation experiments, it is well-recognized that varying segmentation granularities can significantly influence translation quality, particularly in languages like Chinese where clear word boundaries are often absent (Zhao et al., 2013). Therefore, we conduct additional experiments to assess the impact of segmentation granularity on translation performance. Our experiments involve three different segmentation granularities for model input in both Chinese and English datasets: unsegmented, Chinese segmented using the 'Jieba' tool, and Chinese-English segmented using the 'MOSS' tool.

Finally, we compare the performance of our system with commonly used document-level machine translation models. Detailed findings will be presented in the subsequent section.

Translation Prompt	BLEU
TP1	27.92
TP2	27.19
TP3	27.54

Table 2: The results of three candidate translation prompts.

Split the document into k segments	BLEU
k=5	27.73
k=10	27.92
k=15	27.94
k=20	28.08
k=25	27.88
k=30	N/A

Table 3: The results of TP1 with different segment lengths.

4 Results

4.1 Score Analysis

In the discourse-level translation task, we evaluate the performance of three different candidate prompts, as shown in Table 2. Considering these candidate prompts, TP1 yields the highest BLEU score. Therefore, in the subsequent comparative experiments, we consistently employ TP1 as the foundational prompt.

We initially include additional theme information in TP1 based on a suggestion from gpt-3.5-turbo. The theme is related to novels, and we use it to translate the provided sentences from Chinese to English. Surprisingly, the resulting BLEU score is only 27.02, which is even worse than the three base candidate prompts. Consequently, we decide to remove this additional theme information.

For text fragment segmentation, we do experiment with different values of k, including 5, 10, 15, 20, 25, and 30. However, when we set $k = 30$, we encounter errors due to the input being too lengthy for the model to handle. Therefore, we obtain results for the five groups, as shown in Table 3.

We observe that, with the same prompt, varying the length of text segments indeed has an impact on translation performance. When the number of sentences reaches 30 and the token count exceeds 4,096, the system can no longer perform translation. Conversely, when the text length is relatively short ($k = 5$), the model cannot gather enough informa-

Word segmentation granularity	BLEU
unsegmented	27.88
segmented with jieba	**28.16**
segmented with moss	27.53

Table 4: The results of TP1 with different Word segmentation granularity.

Training strategies	BLEU
exp_randinit	21.21
exp_finetune	24.46
exp_mBART	25.26

Table 5: The results of G-transformer with different training modes.

tion, leading to the lowest translation performance. Conversely, overly long text segments ($k = 25$) also weaken performance of the model, potentially introducing noise. Therefore, we choose $k = 20$ as the base for our experiments.

As shown in Table 4, the granularity of text significantly affects the performance of machine translation. Experimental results demonstrate that unsegmented Chinese and English texts are impacted due to the lack of alignment between words, resulting in a slight reduction in translation effectiveness. However, the 'MOSS' segmentation granularity leads to the worst result. We infer that the word segmentation results are too dispersed, making it challenging for the large language model to precisely integrate contextual information for word translation.

Before the widespread use of effective prompts for large-scale models, fine-tuning on pre-trained language models is a common approach to enhance translation performance in specific domains. Therefore, for the comparison experiments, we select a state-of-the-art (SOTA) model designed for document-level machine translation. G-transformer is a straightforward extension of the standard Transformer architecture (Vaswani et al., 2017), using group tags for attention guiding, and introducing locality assumption as an inductive bias to reduce the hypothesis space of the attention from target to source. And we train the G-transformer model using the training corpus provided in the task. This training process involved random initialization, fine-tuning initialization, and fine-tuning on mBART (Liu et al., 2020). The results of these experiments are presented in Table 5.

Comparing the experimental results, it becomes evident that conducting targeted fine-tuning experiments on large language models can enhance machine translation performance. However, it is important to note that this approach falls significantly short of the effectiveness achieved by using prompts on large language models.

4.2 Discourse Analysis

In the context of a document translation (S, T), Lyu et al. (2021) argues that translation consistency should be maintained at the target end if a lexical word w occurs multiple times (two or more times) at the source end.

Due to constraints on time and resources, we conduct manual discourse-level analysis on a limited amount of text. Specific operations are as follows: First, we use a co-reference identification tool (Gardner et al., 2018) to identify all co-reference chains in the target-side documents. We perform data cleaning to extract multiple entity co-reference chains and then compare whether the entity words in the co-reference chains maintain translation consistency.

An example is provided in Table 6. Given that the three candidate prompts exhibit similar discourse characteristics, we choose the large language model gpt-3.5-turbo with prompt TP1 as an example for our analysis. We also introduce the model fine-tuned on the large model mBART for comparison.

Upon observing the result, we notice that even excellent models like ChatGPT may face challenges in addressing certain issues of discourse consistency and coherence. This could be attributed to the extensive training data and the challenge of ensuring coverage of test datasets. On the other hand, fine-tuning strategies, owing to their training on domain-specific data, result in more targeted translations and facilitate the maintenance of translation consistency. This underscores a demand for higher quality document-level translation and could potentially indicate a direction: the need to capture more contextual dependencies.

5 Conclusion

We have presented our experimental study on gpt-3.5-turbo for machine translation, covering translation prompts and robustness. Through careful observation and analysis of the experimental

Source	Reference	Num	Large model with prompt TP1	Num	Finetune on mBART model	Num
佑哥	Brother Assist	12	You Ge	12	You Ge	12
落落	Luo Luo	13	Lulu Luo Luo Luoluo	6 5 2	Luo Luo	13
七月	July	12	July Qiyue	7 6	July	13
烈烈	Lie Lie	19	Lielie Lie Lie	15 4	Lie Lie	14
榜	list	6	board list	4 4	list	5
无誓之剑	Oathless Sword	12	Wu Shi Zhi Jian Oathless Sword	9 2	Oathless Sword	10
韩家公子	Yang Master Han	16	Han Jia Gongzi Han's young master	2 14	Yang Master Han	16

Table 6: The analysis of discourse phenomenon on different translation models.

results, we have noted that the utilization of the large language model with prompts achieves a significant improvement, nearly 3 points higher than the baseline. It even surpasses the currently widely used mBART+fine-tune approach for discourse-level machine translation. We also attempt to enhance translation performance by incorporating in-context information, but this lead to a negative impact. Our future work may include investigating the impact of historical context on translation results and iterative refinement of translation. Simultaneously, we will focus on the recognition and translation of discourse phenomena for large language models.

Acknowledgements

We sincerely thank all the anonymous reviewers for their insightful comments and suggestions to improve the paper. This work was supported by the National Key Research and Development Program of China (2020AAA0108004) and the Key Research and Development Program of Yunnan Province (Grant No. 202203AA080004).

References

Guangsheng Bao, Yue Zhang, Zhiyang Teng, Boxing Chen, and Weihua Luo. 2021. G-transformer for document-level 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 (Volume 1: Long Papers)*, pages 3442–3455, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. *ACL 2018*, page 1.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2023. News summarization and evaluation in the era of gpt-3.

Wenxiang Jiao, Wenxuan Wang, JT Huang, Xing Wang, and ZP Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Xinglin Lyu, Junhui Li, Zhengxian Gong, and Min Zhang. 2021. Encouraging lexical translation consistency for document-level neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3265–3277, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36, Minneapolis, Minnesota. Association for Computational Linguistics.

Hai Zhao, Masao Utiyama, Eiichiro Sumita, and Bao-Liang Lu. 2013. An empirical study on word segmentation for chinese machine translation. In *Computational Linguistics and Intelligent Text Processing*, pages 248–263, Berlin, Heidelberg. Springer Berlin Heidelberg.