Exploring Prompt Engineering with GPT Language Models for Document-Level Machine Translation: Insights and Findings

Yangjian Wu
Lan-Bridge / Sichuan (China)
wuyangjian@lan-bridge.com

Gang Hu
Lan-Bridge / Sichuan (China)
hugang@lan-bridge.com

Abstract

This paper describes Lan-Bridge Translation systems for the WMT 2023 General Translation shared task. We participate in 2 directions: English to and from Chinese. With the emergence of large-scale models, various industries have undergone significant transformations, particularly in the realm of document-level machine translation. This has introduced a novel research paradigm that we have embraced in our participation in the WMT23 competition. Focusing on advancements in models such as GPT-3.5, we have undertaken numerous prompt-based experiments. Our objective is to achieve optimal human evaluation results for document-level machine translation, resulting in our submission of the final outcomes in the general track.

1 Introduction

Recently, large-scale language models, such as GPT-3.5 and GPT-4 (gpt), have emerged as powerful tools in the field of natural language processing. These models have showcased their impressive capabilities in a wide range of tasks, including text generation, question answering, language translation, and more. Language models like GPT-3.5 possess the ability to understand and generate coherent, contextually relevant text, capturing the nuances of language usage and producing high-quality outputs.

In particular, machine translation is an area where these language models have shown tremendous promise. Traditional machine translation models (Yang et al., 2020) used the conventional Transformer architecture (Vaswani et al., 2017) since GPT-3.5 has the potential to revolutionize the translation process by leveraging its massive size and language understanding capabilities. With the advent of large models, the machine translation field has faced new challenges, and utilizing large models for machine translation is a novel attempt.By effectively incorporating prompts and context, GPT-3.5 can produce translations that exhibit fluency, accuracy, and adherence to the source text.

This study focuses on experimenting and evaluating different prompt engineering techniques to further enhance the translation performance of GPT-3.5. By providing more refined and contextually specific prompts, we aim to observe the model's ability to adjust and refine its translations, resulting in improved translation quality. Additionally, we explore the impact of temperature adjustments on the generated translations, allowing us to fine-tune the level of randomness in the output and achieve more deterministic and accurate translations.

Furthermore, we investigate both sentence-level and document-level approaches, examining the effectiveness of GPT-3.5 in handling translations at different granularity levels. These approaches aim to leverage the model's language understanding capabilities to not only produce accurate sentence-level translations but also ensure coherence and consistency at the overall document level.

By delving into these aspects and evaluating the performance of GPT-3.5 in the context of the WMT competition, we aim to contribute to the broader understanding of the capabilities, strengths, and limitations of state-of-the-art language models in the field of machine translation.

The inspiration for this study stems from the outstanding performance exhibited by these large-scale language models, especially in addressing real-world challenges such as major wildfires. GPT-3.5-4k and GPT-3.5-16k, with their increased model capacities, have demonstrated remarkable capabilities in generating high-quality text across various domains. Motivated by these advancements, our study aims to harness the power of these models and explore their potential in the specific domain of machine translation.

By leveraging the robustness and adaptability of GPT-3.5-4k and GPT-3.5-16k, we conduct rigorous

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 166–169
December 6–7, 2023.

experimentation to thoroughly evaluate their translation capabilities. We delve into the nuances of different parameter adjustments, including prompts and temperature, to optimize and enhance the models' performance specifically for translation tasks. By strategically fine-tuning these parameters, we aim to unlock hidden potential and push the boundaries of their translation capabilities.

Real-world challenges, such as major wildfires, require timely and accurate translation of critical information across languages. The effectiveness of machine translation plays a pivotal role in communicating vital updates and ensuring efficient information dissemination during such situations. By investigating the translation capabilities of GPT-3.5-4k and GPT-3.5-16k, we strive to contribute insights that can improve translation efficiency and aid in overcoming language barriers in emergency situations.

With this study, we aim to shed light on the immense potential of large-scale language models, such as GPT-3.5, in addressing real-world challenges through machine translation. By harnessing their capabilities and understanding their performance in various scenarios, we hope to pave the way for more effective and accurate translation systems that can assist in critical situations.

2 Methods

We have designed three prompt schemes:

P1: Translate this sentence from SRC to TGT, do not write any explanations

P2: Translation Request - Sentence-by-Sentence Translation. Language Pair: SRC to TGT. Instructions: 1. Each sentence of the document will be provided individually in the "Original Sentence" section. 2. In the "Translation" section, please provide the corresponding translation for each sentence, considering the context and aiming for faithful translation while minimizing unaligned translations. 3. Avoid including any explanations in the translation. Original Sentence:

P3: Translation Request - Sentence-by-Sentence Translation. Language Pair: SRC to TGT. Instructions: 1. Each sentence of the document will be provided individually in the "Original Sentence" section. 2. In the "Translation" section, please provide the corresponding translation for each sentence, considering the context and aiming for faithful translation while minimizing unaligned translations. 3. Avoid including any explanations in the translation. 4.Please review the translations for verifying that they remain faithful to the original text and provide revised versions accordingly if necessary. If no revisions are needed, provide the translations as they are.

In our study, we conducted several experiments to evaluate the performance of GPT-3.5. The following were the approaches we employed:

- Sentence-to-sentence translation: We used the prompt "Translate this sentence from SRC to TGT, do not write any explanations" to evaluate the model's ability to translate individual sentences accurately.
- Multi-turn dialogue translation: We explored the impact of multi-turn conversations on the performance of GPT-3.5. Using the prompt P1.
- Multi-turn dialogue translation with detailed prompt P3. This experiment aims to test whether GPT-3.5 has the ability to get faithful translations while minimizing unaligned translations.
- Comparison between GPT-3.5-4k and GPT-3.5-16k: We performed separate experiments using both GPT-3.5-4k and GPT-3.5-16k models to observe any differences in translation abilities between the two.
- Adjusting temperature parameter: We varied the temperature parameter (0, 0.3, 0.7) to examine its impact on the translation quality. Changing the temperature can control the randomness of the generated translations.
- Incorporating fake CoT prompt P3. This experiment aims to test whether GPT-3.5 has the ability to automatically reflect and optimize its translations.

3 Result

We conduct experiments to quantify the impact of each component in our system. The evaluation conduct on test set on wmt22 using SacreBLEU (Post, 2018) and COMET (Stewart et al., 2020).

As shown in Table 1, here are the conclusions based on your experimental results:

- From the first and second experiment results, it can be concluded that the performance of GPT-3.5 in multi-turn dialogue is better than

language pair	Prompt	Multi-turn	T	Model	Bleu-A	Bleu-B	Chrf-A	Chrf-B	Comet-A	Comet-B
zh-en	P1	false	0	GPT-3.5-4k	26.6	20.0	57.4	52.5	52.7	43.5
zh-en	P1	true	0	GPT-3.5-4k	27.7	20.7	58.4	53.2	55.8	46.7
zh-en	P3	true	0	GPT-3.5-16k	23.4	18.0	54.4	50.2	54.9	46.1
en-zh	P1	true	0	GPT-3.5-4k	45.7	53.9	41.1	48.5	63.4	71.2
en-zh	P2	true	0	GPT-3.5-4k	44.2	51.4	39.9	46.0	62.1	70.6
en-zh	P2	true	0.7	GPT-3.5-4k	42.8	49.3	38.4	44.1	61.7	68.7
en-zh	P2	true	0.7	GPT-3.5-16k	42.7	49.3	38.3	44.8	63.2	71.1
en-zh	P2	true	0.3	GPT-3.5-16k	44.4	51.5	39.9	46.3	63.8	71.2

Table 1: Bleu/Chrf/Comet score on wmt22 test set. The COMET scores are calculated with the model wmt20-comet-da, the ChrF scores are calculated using all available references and SacreBLEU signature is the default settings. Scores are multiplied by 100. T represents Temperature

single-turn translation. This indicates that context can help improve the translation quality of GPT-3.5 by providing additional prompts.

- Comparing the results of the third experiment with the fourth experiment, it is concluded that the performance of P2 is worse. This suggests that GPT-3.5 does not fully understand the given prompt, which results in difficulty in generating accurate translations.

- Comparing the results of the fourth, fifth, and seventh experiments, it is concluded that lower temperature values yield better translation results. This indicates that reducing temperature parameter leads to more deterministic and high-quality translations.

- Comparing the results of the fifth and sixth experiments, it is concluded that GPT-3.5-16k performs better in translation than GPT-3.5-4k.

- Comparing the results of the seventh experiment with previous results, it is concluded that P3 performs the worst. Additionally, observing the actual revised results, it can be noted that GPT-3.5-16k rarely modifies its translations, indicating that without specific and clear instructions, it is unable to make effective modifications to its own translations.

Based on our previous results, we have chosen GPT-3.5-16k as the final model for our submission. For the WMT23 en-zh/zh-en track, we set the temperature to 0 and utilized P1 as the prompt. Adopting a multi-turn dialogue approach, we submitted our final results with the systerm name "LanBridgeMT". Figure 1 and Figure 2 show the results of our systerm. [1] Additionally, for other language pairs in the general WMT competition, we opted to submit the results generated by our LanMT (Han et al., 2022) engine. This decision was made to assess the engine's performance and determine its scoring capabilities directly in the online evaluation environment.

By taking these approaches, we aim to showcase the effectiveness of GPT-3.5 and demonstrate the performance of our LanMT engine in the respective WMT tracks. These submissions reflect our overarching goal of participating in and contributing to the advancement of machine translation research and development.

4 Conclusion

In this study, we evaluated the translation performance of GPT-3.5 using various experimental approaches. Our findings indicate that incorporating multi-turn dialogue prompts improves the translation quality of GPT-3.5, highlighting the importance of context in guiding the model's translations. Furthermore, we observed that GPT-3.5-16k, compared to GPT-3.5-4k, demonstrates superior translation capabilities in commit scores, indicating its enhanced ability to understand and fulfill user instructions. However, there are marginal differences in the other two metrics, BLEU and ChrF. Additionally, we found that lower temperature values result in improved translation quality, indicating the usefulness of controlling randomness in the generated translations. However, it is important to note that GPT-3.5 may struggle with understanding ambiguous prompts and lacks the ability to autonomously adjust and optimize its translations without explicit instructions. These findings contribute to our under-

[1] https://github.com/wmt-conference/wmt23-news-systems

System	COMET
HW-TSC	82.8
ONLINE-B	82.7
Yishu	82.7
GPT4-5shot	81.6
Lan-BridgeMT	81.2
ONLINE-G	80.9
ONLINE-Y	80.6
ONLINE-A	80.3
ZengHuiMT	79.6
ONLINE-W	79.3
IOL_Research	79.2
ONLINE-M	77.7
NLLB_MBR_BLEU	76.8
ANVITA	76.6
NLLB_Greedy	76.4

System	chrF
HW-TSC	57.5
ONLINE-B	57.5
Yishu	57.4
ZengHuiMT	54.6
ONLINE-G	53.9
ONLINE-A	53.4
GPT4-5shot	53.1
Lan-BridgeMT	53.1
ONLINE-W	52.5
IOL_Research	52.4
ONLINE-Y	52.3
ONLINE-M	49.7
ANVITA	47.1
NLLB_Greedy	46.1
NLLB_MBR_BLEU	45.8

System	BLEU
HW-TSC	33.6
ONLINE-B	33.5
Yishu	33.4
ONLINE-A	28.3
Lan-BridgeMT	27.3
IOL_Research	27.2
ZengHuiMT	27.0
GPT4-5shot	26.8
ONLINE-G	26.6
ONLINE-W	26.4
ONLINE-Y	25.0
ONLINE-M	23.5
ANVITA	21.8
NLLB_Greedy	20.5
NLLB_MBR_BLEU	19.8

Figure 1: Score for zh-en translation task

System	COMET
ONLINE-B	88.1
Yishu	88.1
HW-TSC	87.3
GPT4-5shot	87.1
ONLINE-W	86.8
Lan-BridgeMT	86.6
ONLINE-Y	86.5
ONLINE-A	86.2
IOL_Research	85.3
ZengHuiMT	84.3
ONLINE-M	84.2
ONLINE-G	83.8
NLLB_Greedy	75.7
ANVITA	75.6
NLLB_MBR_BLEU	71.5

System	chrF
HW-TSC	53.8
Yishu	53.0
ONLINE-B	52.9
ONLINE-A	52.8
IOL_Research	51.9
ONLINE-M	50.6
ONLINE-Y	49.8
ONLINE-G	49.4
ONLINE-W	47.3
ZengHuiMT	47.0
Lan-BridgeMT	46.8
GPT4-5shot	46.5
ANVITA	36.9
NLLB_Greedy	26.3
NLLB_MBR_BLEU	21.1

System	BLEU
HW-TSC	58.6
ONLINE-A	58.5
Yishu	57.6
ONLINE-B	57.5
IOL_Research	56.9
ONLINE-M	54.9
ONLINE-Y	54.2
ONLINE-G	54.1
ZengHuiMT	52.9
ONLINE-W	52.1
Lan-BridgeMT	50.2
GPT4-5shot	49.6
ANVITA	38.9
NLLB_Greedy	27.4
NLLB_MBR_BLEU	19.1

Figure 2: Score for en-zh translation task

standing of the strengths and limitations of GPT-3.5 in translation tasks, emphasizing the need for precise prompts to achieve optimal translation results.

References

Gpt-4 technical report.

Bing Han, Yangjian Wu, Gang Hu, and Qiulin Chen. 2022. Lan-bridge MT's participation in the WMT 2022 general translation shared task. In Proceedings of the Seventh Conference on Machine Translation (WMT), pages 268–274, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Craig Stewart, Ricardo Rei, Catarina Farinha, and Alon Lavie. 2020. COMET - deploying a new state-of-the-art MT evaluation metric in production. In Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track), pages 78–109, Virtual. Association for Machine Translation in the Americas.

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.

Shuoheng Yang, Yuxin Wang, and Xiaowen Chu. 2020. A survey of deep learning techniques for neural machine translation. arXiv preprint arXiv:2002.07526.