Hou, Xinyu
International University SDI Munich
Supervisor – Mayer, Felix, PhD, Professor
https://doi.org/10.33989/pnpu.791.c3287

Bridging Languages and LLMs in Translation: A Study on Prompt Engineering for Customized LLM Model Using the COSTAR Framework

1. Introduction

"AI will continue to get way more capable and will become ubiquitous as time goes on. People are using it to create amazing things. If we could see what each of us can do 10 or 20 years in the future, it would astonish us today." (Altman, 2024)

This quote comes from OpenAI CEO Sam Altman and reflects his optimistic view on the future of Artificial Intelligence (AI). The increasing capabilities of this technology have also made the application of large language models a hot topic. Particularly in machine translation, models like ChatGPT have demonstrated exceptional performance in translation tasks, becoming essential tools for professional translators. This study aims to enhance the quality of machine translation, especially for specialized texts such as patent documents, through prompt optimization and terminology extraction using the COSTAR framework. Special attention is given to the comparison between a customized GPT model, adapted with suitable prompts following the CO-STAR framework and targeted terminology optimization, and the regular ChatGPT-4 model. The influence of terminology and specific prompts on specialized translations is also considered to increase the efficiency and precision of machine translation.

2. Research Background

Neural Machine Translation (NMT) is the most advanced approach in machine translation and is based on deep neural networks capable of recognizing complex patterns in large datasets and improving translation through "training". A groundbreaking model in NMT is the Transformer, which utilizes the self-attention mechanism. Due to these features, the Transformer has proven particularly suitable for translating complex sentences and longer texts. Although NMT has made impressive progress and has become the standard for machine translation, challenges remain, such as high computational demands and a tendency to make errors with longer sentences (cf. Pérez-Ortiz et al., 2022, pp. 141-164).

ChatGPT is a conversational AI developed by OpenAI, based on the GPT (Generative Pre-trained Transformer) architecture. The model benefits from extensive natural language processing and utilizes data from diverse sources to broaden the range of its responses. Through the use of reinforcement learning, the accuracy and relevance of ChatGPT's responses have been significantly enhanced (cf. Thakur, Barker, & Pathan, 2024, p. 89).

The goal of this study is to improve ChatGPT's performance in patent translation by customizing prompts using the COSTAR framework and optimizing terminology. Specifically, the research compares the translation quality of a

customized GPT model with that of the standard ChatGPT-4 model, focusing on terminology consistency, fluency, and accuracy.

3. Methodology: COSTAR Framework and Prompt Optimization

The COSTAR framework is a systematic approach to prompt design, aimed at improving LLM outputs by clarifying the context, objective, style, tone, audience, and response format. Developed by GovTech Singapore's Data Science & AI team, this framework is widely used for complex tasks. By systematically applying this approach, more relevant and precise results can be achieved, as the framework ensures that all relevant factors are clearly and comprehensively integrated into the prompt (cf. Teo, 2023). The key components of COSTAR are:

Context (C): Provide background information to help the model understand the scenario.
Objective (O): Define the task goal to ensure the model focuses on the desired outcome.
Style (S): Specify the writing style (e.g., professional, formal, or humorous).
Tone (T): Set the tone to align with the required sentiment or emotional context.
Audience (A): Identify the target audience (e.g., experts, beginners, or children).
Response (R): Define the output format (e.g., list, JSON, or report).

In this study, the COSTAR framework was used to design prompts for patent document translation. Below is an example of a COSTAR-based prompt:

"Context: You are a translation expert specializing in translating German patent documents into Chinese. Your task is to ensure accuracy, terminology, linguistic norms, style, and formatting so that the translation meets the requirements of patent documents. The translation should be created in accordance with the TQA2024 quality standards, taking into account the error classification and severity levels of the DGT guidelines. Key error categories include accuracy (ACCY), terminology (TERM), linguistic norms (LNORM), style (STJOB and STGEN), and design (DSGN). Objective: Translate the following patent text to comply with the TQA2024 quality standards for patent documents. Ensure that all technical terms are correctly translated and that the text meets the formal linguistic and stylistic requirements of patents. Style: Professional and precise to ensure the translation meets the formal requirements of patents. Tone: Clear and informative, with a focus on the accurate representation of technical and legal content. Target Audience: Patent attorneys, engineers, and professionals involved in the translation and application of patents. Response: Translate the following text, ensuring all technical terms are correctly translated. Provide explanations for ambiguous terms or specialized expressions if necessary. Ensure terminology consistency and formatting compliance with TQA2024. For each identified error, specify the error type and severity level. Ensure the translation achieves a quality score of at least 80% according to TQA2024 guidelines and calculate a quality rating (0% to 100%) based on errors and the DGT document guidelines."

By using this structured prompt design, the model can better understand the task requirements and generate higher-quality translations.

Customized GPT link: https://chatgpt.com/g/g-673d24115cc881918f1b121727bc87dd-patent-translator-pro

4. Experimental Design and Implementation

The experiment consisted of three phases: preparation, production, and evaluation.

4.1 Preparation Phase

The preparation phase focuses on data collection, GPT model configuration, prompt creation, and general preparatory tasks. The selected patent documents were collected from the German Patent and Trademark Office (DPMA), and corresponding Chinese reference translations were obtained from Google Patents. Terminology

extraction was performed using Trados Studio, resulting in a list of 17 specialized terms, which were integrated into the GPT model to ensure terminology consistency. In the process of configuring the GPT model, I developed a model specifically tailored for translating patent applications, based on ChatGPT-4. This model was named “Patent Translator Pro”. The customization was carried out using the previously defined prompts created according to the COSTAR framework. Additionally, a terminology list extracted earlier was integrated into the model to enhance the accuracy of terminology in translations. Furthermore, I uploaded the TQA 2024 EU standard into the system to ensure that the translations from the customized GPT model meet relevant quality requirements and achieve high-quality patent application translations.

4.2 Production Phase

The translation process includes revising the reference translation and using ChatGPT-4 and the customized GPT model (Patent Translator Pro) to translate the original German text. The goal is to summarize the translation results from different models to determine which model delivers the best translation version.

During the experiment, I observed that the Chinese translation of the officially published patent documents was of lower quality. Issues such as omissions, mistranslations, terminology problems, and issues with word order and grammar significantly affected the evaluation of translation quality, leading to scores that did not reflect the actual quality. For this reason, my second supervisor and I revised and improved the officially published reference translation. The revised version is used as the reference translation in the subsequent evaluation of translation quality.

Results of the Customized GPT Model (Patent Translator Pro):

[Translated Chinese text provided in reference]

Results of GPT-4:

[Translated Chinese text provided in reference]

4.3 Evaluation Phase

In the evaluation phase, the translation outputs were assessed using the automatic metrics such as BLEU, METEOR, BERTscore, and COMET. These metrics provide an objective assessment of translation performance based on various criteria. BLEU and METEOR are used to measure alignment with reference translations, while BERTScore analyzes semantic similarities between machine translations and references. Additionally, COMET is utilized to enable a more comprehensive evaluation of translation quality, incorporating contextual information. These metrics complement each other and provide a nuanced view of translation quality, ensuring a thorough evaluation. The goal is to obtain a quantitative analysis of translation quality, which serves as a basis for improving the models.

To automatically evaluate the quality of machine translations, I combined PyCharm and Anaconda and designed programs in different environments to develop an automated evaluation process based on four evaluation metrics. This process includes inputting multiple translation examples, calculating quality results, and comparing and visualizing them.

Results:

The customized GPT model outperforms ChatGPT-4 across multiple evaluation metrics, achieving a BLEU score of 0.74 compared to ChatGPT-4's 0.51, a METEOR score of 0.88 versus 0.72, a BERTScore of 0.9338 against 0.8602, and a COMET score of 0.8150 compared to 0.7311.

5. Results and Analysis

The results demonstrated that the customized GPT outperformed the standard ChatGPT-4 model in terms of terminology consistency and fluency. Specifically, the model excelled in handling complex terminology and long sentence structures in patent documents. The outstanding performance of the customized model can be attributed to the COSTAR prompts specifically developed for patent literature translation. These prompts, combined with an integrated terminology list, enabled the model to better handle the unique requirements of patent texts, use precise terminology, and improve translation quality. Additionally, the author created a table summarizing the scores of both models under the various evaluation methods, further highlighting the advantages of the customized GPT model.

The results highlight the potential of advanced prompt engineering and model customization to improve machine translation quality, particularly in specialized domains such as patent translation. The customized GPT model not only demonstrates technical precision but also linguistic fluency, making it a promising tool for professional translation tasks in this field. By leveraging tailored prompts, domain-specific terminology, and iterative optimization, the customized model effectively addresses the unique challenges of patent translation, setting a new benchmark for machine translation in specialized contexts.

6. Conclusion

This study has experimentally proven that a specially customized GPT model, optimized through carefully designed prompts for translating specialized patent documents, delivers significantly higher translation quality compared to the general GPT model. Although the translation performance of ChatGPT-4 is already very good, the experimental results demonstrate that a customized GPT model provides significantly better outcomes. Especially in the field of patent translation, the use of an LLM not only saves time and costs but also delivers high-quality and satisfactory results. The adaptation of prompts for the GPT model based on the CO-STAR Framework has successfully bridged the gap between machine translation and artificial intelligence. This provides valuable support for future translators, especially in the translation of highly specialized texts and patent documents, and opens up new perspectives for efficiency and precision in the translation industry.

In summary, research in the field of LLMs will continue to progress intensively, particularly in terms of model optimization and terminology management. The potential of new emergent capabilities in large language models could further revolutionize their application in translation, especially for complex patent documents. Well-designed prompt engineering will continue to play a central role in improving translation quality in the future.

References

European Union (2024). *Translation Quality Evaluation Info Pack for External Contractors*. https://eur-lex.europa.eu/eli/dec/2011/833/oj. [15.08.2024].
Github. (2024). *COMET*. https://github.com/Unbabel/COMET [12.10.2024].
Google Patents. (2019). *DE 102011086742B4*.
https://patents.google.com/patent/US20240341407A1/en?q=(eletronic+german)&oq=el
etronic+german&sort=new&page=1&clustered=true#citedBy [20.08.2024].
Hou, X.Y. Google Drive. (2024). *Masterarbeit Dokumente*.
https://drive.google.com/drive/folders/1DbrHH8SdQM2d6e4om0yy_AhfeFQYviEq. [27.02.2025].
MIT Sloan. (2024). *Sam Altman believes AI will change the world*. https://mitsloan.mit.edu/ideas-made-to-matter/sam-altman-believes-ai-will-change-world-and-everything-else [12.07.2024].
Pérez-Ortiz, Juan Antonio, Mikel L. Forcada & Felipe Sánchez-Martínez (2022). *How neural machine translation works*. In Dorothy Kenny (ed*.). Machine translation for everyone: Empowering users in the age of artificial intelligence*, 141–164. Berlin: Language Science Press. DOI: 10.5281/zenodo.6760020.
Teo, S. (2023). *How I won Singapore's GPT-4 prompt engineering competition*. *Towards Data Science.* https://towardsdatascience.com/how-i-won-singapores-gpt-4-prompt-engineering-competition-34c195a93d41 [12.08.2024].
Thakur, K., Barker, H.G., & Khan Pathan, A.-S. (2024). *Artificial Intelligence and Large Language Models: An Introduction to the Technological Future* (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781003474173.

Xinyu Hou
International University SDI Munich

Bridging Languages and LLMs in Translation: A Study on Prompt Engineering for Customized LLM Model Using the COSTAR Framework

This article focuses on the implementation of a specialized AI assistant based on the ChatGPT language model for translation tasks. By employing specific prompts and conducting translation experiments, the effectiveness of these methods will be empirically analyzed. The results can improve the efficiency and quality of translations in specific industrial sectors, thereby bridging the gap between different languages and industries.

***Key words**: large language models, machine translation, prompt engineering, ChatGPT, translation quality evaluation.*

Li, Xintian
International University SDI München
Supervisor – Mayer, Felix, PhD, Professor

https://doi.org/10.33989/pnpu.791.c3288

A Comparative Experiment of Traditional Tools and LLM Tools in Terminology Extraction: Sketch Engine and ChatGPT as Examples

Terminology plays a pivotal role in both our everyday and professional lives. Identifying widely accepted expressions marks the beginning of translation and interpreting work, while reviewing and refining terminology completes the process. With the rise of various Large Language Model (LLM) tools and the widespread speculation that LLMs may replace many translators and interpreters in the market, newcomers to the field may feel uncertain about how to navigate the constantly evolving landscape of the profession. Given the undeniable importance of