System for Improving Number Translation Using LLMs

Kenneth Chen

https://github.com/kchen190053/CS6375Project

University of Texas at Dallas

CS6375 Wei Yang

*Abstract*— *Modern large language models (LLMs) have improved drastically in both capability and accessibility within the past few years, however they are not without issues. One issue is that LLMs frequently falter when machine translating numerical data. This study evaluates four different LLM-based machine translation services’ performance when translating between Mandarin Chinese and English using In-context learning, Chain of thought, and TEaR framework methods. Afterwards*

Keywords—LLM, Machine Translation, Numerical Translation, Translation

# Introduction

In today’s interconnected world, encountering a need to understand foreign languages is commonplace. Many people utilize machine translation services for their convenience and accessibility. In STEM fields numerical translation is particularly important. Numerical translation requires high precision when compared with text translation, as an incorrect decimal point or digit may result in large deviations from intended results. A mistranslation may turn a thousand into a hundred and result in a business ordering too few supplies. Ensuring reliable and precise numerical translation is thus imperative for business transactions and scientific research collaboration.

Numerical translation using LLMs faces many challenges as different language often treat numbers in unique ways, such as with different units or separators. Standard metrics such as BLEU or COMET also do not capture the severity of numerical errors. In this paper, we utilize several open source LLMs and MTL services and evaluate their reliability and precision for translating numerical data between Chinese and English. We also analyze the effectiveness of the TEaR framework [5] for numerical translation.

# Background

Translation accounts for a small portion of work done by LLMs, and numerical translation for LLMs is an even narrower field of research. As such, research in this specific area is scarce. One notable study in this area is [1], assessing Neural Machine Translation for numerical translation. The study found that open-source NMT models do not have 100% accuracy in numerical translation. The languages they tested had ranges of 61% to 96% for accuracy; a chance of error too large for reliable and trustworthy use for real world projects. A study by Sweta et al. [2] found that using In-context learning improved translation performance. By conditioning the language model with in-context parallel examples, the model can more effectively translate text and also mimic the style of in-context examples in addition to being used for template-based translation, however, the study did not directly tackle numerical translation. A separate study [3] found that using LLMs as automatic post-editors instead of translators resulted in higher translation performance. Once again, this technique was not specifically utilized for numerical translation. Finally, common machine translation metrics do not capture the severity of translation errors. BLEU treats words/numbers as discreet tokens, so a minor rounding error in translation cannot be distinguished from major errors like an incorrect magnitude. COMET relies on pre-trained multilingual language models to assess the semantic similarity between the source, translation, and reference, and its performance is tied to the data it was trained on, and if that data didn't adequately expose the model to the nuances and importance of numerical accuracy in various contexts, it might not learn to penalize numerical errors appropriately. In addition, COMET’s holistic score does not identify or categorize numerical errors, and it may not reflect the severity of numerical translation errors if surrounding context remains largely the same.

# Related Works

In a study done by Tang et al. [4], work is done to evaluate the effectiveness of machine translation with LLMs using ten different criteria for accurate numerical translation, which showed that in some cases LLM numerical translation accuracy is unacceptable, such 78% accuracy when translating large numbers between Chinese and English. This paper seeks to evaluate other LLM models that [4] did not analyze and utilize other strategies to improve translation reliability. Feng et al [5] introduced a framework named TEaR for the purpose of improving LLM machine translation and demonstrated its effectiveness for improving the translation of text. It did not apply it for the specific task of numerical machine translation however, and this paper seeks to test its effectiveness on numerical translation.

# Formal Problem Definition

The inaccurate translation of numbers can lead to errors in scientific and business applications, not to mention confusion in literary contexts. LLMs can be used to streamline work flow, such as summarizing a business or scientific document from another language, or for annotation purposes. However, this method cannot be used if the reliability of the LLM for numerical translation is unsatisfactory. This paper seeks to address this issue by utilizing various techniques and frameworks with different models for the purpose of Chinese to English and English to Chinese translation. Below are scenarios in which translation errors may be found when translating between the two languages.

**Units:** Converting units from one language to another, such as translating “万” to “ten thousand”. Some units do not have exact equivalents in the other language.

**Separators:** Accurate translation ofnumberswhile ensuring the correct placement of commas and decimal points. Some numbers should not have separators, such as phone numbers and ID numbers. Translation should not add additional separators such as hyphens or commas when not applicable.

**Numerals:** Translation of word into numerals, such as from “二万二千四十五” to “22,045” and vice versa, from numbers to word form.

**Literary:** Phrases that involve numbers may have issues during translation depending on how literally the phrase is translated. The unit “万” is particularly noteworthy due to there not being an equivalent unit in English, in addition to “万” in some contexts is used to represent “countless” or “myriad” rather than a set number. “万事如意” means “best wishes”, or “all wishes will be fulfilled”, using the “万” character as “a lot” instead of as ten thousand.

# Approach

This paper will test the methods on models Gemini 2.0, GPT-4o, DeepSeek R1, and Llama 2 7B. The methods we will use are In-context-learning (ICL), Chain-of-Thought (COT), and TEaR[5]. In-context learning is essentially a method of helping models learn how to perform tasks quickly by including a list of input-output pairs within a prompt. The prompt includes training examples with a correct output matched to an input, allowing models to locate a previously learned concept within the model’s large training data and improve task learning performance. Chain of thought prompting enables models to perform more complex tasks using intermediate reasoning steps. Even one intermediate example step being included in the prompt can improve model performance. The TEaR framework is broken down into three steps: Translate, Estimate, and Refine. The translate step utilizes zero-shot prompting, just directing a model to perform translation. The Estimate step prompts the model to follow a specific typology to emulate a human annotator’s evaluation of the translation, identifying translation errors such as critical factual errors or stylistic awkwardness. If no errors are detected, the translation performed in the Translate step is designated as the final output. If errors are found, the model is asked to perform the Refine step, retranslating the original text with the errors detected in mind within the prompt, allowing for LLM self-refinement. In-context learning and chain-of-thought methods have previously been applied to numerical translation, however previous work have not tackled these specific models. TEaR framework has not been applied numerical translation, and ICL and COT can serve as comparison points to evaluate how well the TEaR framework performs when attempting to improve numerical translation. Passages were selected from business documents of Chinese companies to use as texts to be translated by the models. The documents used were the 2023 and 2024 quarterly earning reports of Chinese companies BYD and Xiaomi. Using passages selected from actual Mandarin Chinese business documents as the subject means that the results of this study should be applicable to other business documents in similar “business speech” style.

# Implementation

Examples for the prompts used for each method are as follows: The example passage being translated is: “互聯網服務分部收入由2023年第一季度的人民幣70億元增加14.5%至2024年第一季度的 人民幣80億元，主要是由於廣告業務收入增加，惟部分被遊戲業務收入減少所抵銷。境 外互聯網用戶數的持續增長使境外互聯網服務收入由2023年第一季度的人民幣18億元增 加39.0%至2024年第一季度的人民幣25億元。”

Using ICL, the prompt would be:

‘You are a professional Chinese to English translator. Translate the [Source] sentence into English based on the following unit conversions.

100 million = 1 亿

10 million = 1000 万

1 million = 100 万

100 thousand = 10 万

10 thousand = 1 万

1 thousand = 1 千

1 hundred = 1 百

[Source] = 互聯網服務分部收入由2023年第一季度的人民幣70億元增加14.5%至2024年第一季度的 人民幣80億元，主要是由於廣告業務收入增加，惟部分被遊戲業務收入減少所抵銷。境 外互聯網用戶數的持續增長使境外互聯網服務收入由2023年第一季度的人民幣18億元增 加39.0%至2024年第一季度的人民幣25億元。’

Using COT, the prompt would be:

‘You are a professional Chinese to English translator. Translate the [Source] sentence into English step by step. Pay close attention to unit conversions between Chinese and English. Translate numerical parts first, then the sentence.

[Source] = 互聯網服務分部收入由2023年第一季度的人民幣70億元增加14.5%至2024年第一季度的 人民幣80億元，主要是由於廣告業務收入增加，惟部分被遊戲業務收入減少所抵銷。境 外互聯網用戶數的持續增長使境外互聯網服務收入由2023年第一季度的人民幣18億元增 加39.0%至2024年第一季度的人民幣25億元。’

Using the TEaR framework, the prompt for step 1: Translate would be:

‘Please provide the English translation for the Chinese source.

[Source] = 互聯網服務分部收入由2023年第一季度的人民幣70億元增加14.5%至2024年第一季度的 人民幣80億元，主要是由於廣告業務收入增加，惟部分被遊戲業務收入減少所抵銷。境 外互聯網用戶數的持續增長使境外互聯網服務收入由2023年第一季度的人民幣18億元增 加39.0%至2024年第一季度的人民幣25億元。’

The result of the previous prompt would replace the [Translation] text of the Estimate step (the MQM typology is taken from [5]):

‘Based on the Chinese [Source] and English translation, identify error types in the translation and classify them. The categories of errors are accuracy (addition, mistranslation, omission, untranslated text), fluency (character encoding, grammar, inconsistency, punctuation, register, spelling), locale convention (currency, date, name, telephone, or time format) style (awkward), terminology (inappropriate for context, inconsistent use), non-translation, other, or no-error. Each error is classified as one of three categories: critical, major, and minor.  Critical errors inhibit comprehension of the text. Major errors disrupt the flow, but what the text is trying to say is still understandable. Minor errors are technical errors but do not disrupt the flow or hinder comprehension.

Chinese [Source] = 互聯網服務分部收入由2023年第一季度的人民幣70億元增加14.5%至2024年第一季度的 人民幣80億元，主要是由於廣告業務收入增加，惟部分被遊戲業務收入減少所抵銷。境 外互聯網用戶數的持續增長使境外互聯網服務收入由2023年第一季度的人民幣18億元增 加39.0%至2024年第一季度的人民幣25億元。

English Translation = [Translation]’

If the Estimate step does not produce any annotations, the output of the Translate step would be accepted as the output. If annotations are produced, the annotations will replace the [Annotations] text within the prompt for step 3: Refine:

‘I’m not satisfied with this translation because of these defects:

[Annotations]

Upon reviewing this feedback, please retranslate the [Source] sentence.

[Source] = 互聯網服務分部收入由2023年第一季度的人民幣70億元增加14.5%至2024年第一季度的 人民幣80億元，主要是由於廣告業務收入增加，惟部分被遊戲業務收入減少所抵銷。境 外互聯網用戶數的持續增長使境外互聯網服務收入由2023年第一季度的人民幣18億元增 加39.0%至2024年第一季度的人民幣25億元。’

Zero-shot prompting was used to establish a baseline accuracy for each model, followed by performing each of the three methods for each of the four models for each text passage.

# Evaluation

Upon receiving the output of each method on a passage, the number of correct numerical translations was compared to the total number of numerical translations. The results are shown below in Table I.

Table I

Accuracy of Different Methods across Models

A screenshot of a calculator

AI-generated content may be incorrect.

Utilizing methods showed slight improvements in numerical translation accuracy. The baseline translation quality was already high for three of the four models. This may be due to the high existing amount of Chinese training data available. [1] was performed in 2021, and with less popular languages such as Nepalese, which may be why the baseline translation quality was higher. The most common error observed was that of unit translation. The “亿”(hundred million) and “万” (ten thousand) characters were frequently misinterpreted into English units such as billion or million without regard to their differences. Separators and phrases did not result in any errors. This may be due to the documents being written formally in business speech, resulting in more standardized formats which may have facilitated more effective translation. The Llama 2 7b performed the worst by far, never rising above 88% even after applying the three methods. Utilizing TEaR showed an improvement over baseline, however it was less than the improvement caused by utilizing ICL and COT methods.

# Discussion

The biggest limitation of this work is that it only tackles numerical translations from Mandarin Chinese to English, both languages which have a large amounts of training data which may have made the baseline translation quality high. Another major limitation is the choice of passages. The passages came from 2023 and 2024 quarterly reports. The release times of these document are pertinent, as the 2023 reports may have been included in the training data for the GPT, Gemini, and DeepSeek models, which may be why the baseline accuracy for these models were so high, while the Llama 2 7b model’s baseline performance was much lower, as the Llama 2 7b’s training data cutoff did not include these documents. Future work may include testing passages from 2025 or creating passages with similar syntax to but not exact copies of those documents.

# Conclusion

In this paper we have evaluated the effectiveness of In-context-learning, Chain-of-thought prompting, and the TEaR framework for the purposes of numerical machine translation of business documents from Mandarin Chinese to English. Utilizing the methods showed slight improvements over baseline zero-shot prompting for translation. ICL and COT were more effective than utilizing the TEaR framework in improving accuracy. Of the four models tested, Llama 2 7b performed the worst. However, the baseline accuracy was already quite high, perhaps due to poor selection of passages to be translated or the large volume of training data available for the two languages. Further work may include testing other languages, or selecting other passages to translate with more recent publication dates.

##### References

1. Jun Wang, Chang Xu, Francisco Guzm´an, Ahmed El-Kishky, Benjamin I. P. Rubinstein, and Trevor Cohn, “As easy as 1, 2, 3: Behavioural testing of NMT systems for numerical translation,” in Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021, 2021, vol. ACL/IJCNLP 2021 of Findings of ACL, pp. 4711–4717.
2. Sai Koneru, Miriam Exel, Matthias Huck, and Jan Niehues, “Contextual refinement of translations: Large language models for sentence and document-level post-editing,” in Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024. 2024, pp. 2711 2725, Association for Computational Linguistics
3. Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad, “In-context examples selection for machine translation,” in Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023. 2023, pp. 8857–8873, Association for Computational Linguistics
4. Tang, Wei, Jiawei Yu, Yuang Li, Yanqing Zhao, Weidong Zhang, Wei Feng, Min Zhang, and Hao Yang. "Investigating Numerical Translation with Large Language Models." *arXiv preprint arXiv:2501.04927* (2025).
5. Feng, Z., Zhang, Y., Li, H., Wu, B., Liao, J., Liu, W., ... & Liu, Z. (2024). Tear: Improving llm-based machine translation with systematic self-refinement. *arXiv preprint arXiv:2402.16379*.
6. Donthi, S., Spencer, M., Patel, O., Doh, J. Y., Rodan, E., Zhu, K., & O’Brien, S. (2025, March). Improving LLM abilities in idiomatic translation. In *Future of Information and Communication Conference* (pp. 361-375). Cham: Springer Nature Switzerland.