International Journal of Education and Humanities (IJEH), 4(2) 2024:121-128
http://i-jeh.com/index.php/ijeh/index
E-ISSN: 2798-5768

Comparison of Translation Quality between Large Language Models and Neural Machine Translation Systems: A Case Study of Chinese-English Language Pair

Xinchen Li[1]

Abstract

A number of Neural Machine Translation (NMT) systems have already demonstrated their strength to undertake various translation tasks which are not too demanding. However, the incredible advancement of AI technology in recent years has endowed Large Language Models (LLMs) with great potential, so we may imagine that they may even do better than NMT working as translators. To figure out whether LLMs have better performance than NMT in translation, and how genres and translation directions may influence translation quality, this article chose two LLMs, namely, ChatGPT 3.5 and Wenxin Yiyan, or ERNIE Bot 3.5, and one NMT system, namely, DeepL, to test and compare their performance in Chinese-English translation, employing a quantitative method including BLEU scoring and SPSS analysis. The results show that there is no significant improvement in these LLMs' translation quality compared with the NMT system, and all the chosen systems tend to perform better in non-literary translation than in literary translation and produce TTs of higher quality in Chinese-English translation than in English-Chinese translation.

Keywords: *Large language model, Neural machine translation, Translation quality assessment, Chinese-English translation, Comparative study*

A. Introduction

There have already existed a large number of Neural Machine Translation (NMT) systems with acceptable capability to bear translation tasks, and the advancement of artificial intelligence in recent years has given a boost to the development of these systems to make them behave more like a human translator. The most outstanding achievement of AI is ChatGPT launched in 2022, an intelligent chatting machine. ChatGPT is the abbreviation for Chat Generative Pre-Trained Transformer. Developed by OpenAI, it was trained to respond to users' requests wisely in great detail, including accomplishing translation tasks, and is much more capable than any other previous AI of generating coherent and context-aware text (Hendy et al., 2023). Therefore, chances are that, in the field of translation, ChatGPT may perform better than NMT systems.

Besides ChatGPT, there is another large language model called Wenxin Yiyan, or ERNIE Bot, which stands for Enhanced Representation through Knowledge Integration, and it is developed by Baidu, a China-based high-tech company, with features similar to ChatGPT. The development of ERNIE Bot begun in 2019, and was released on March 17 in 2023. According to its self-description, ERNIE Bot can talk and interact with people, answer questions, assist in creation, and help people acquire information, knowledge and inspiration efficiently and conveniently. It can be hypothesised that both of ChatGPT and ERNIE Bot may do a better job

[1]School of Arts, English and Languages, Queen's University Belfast, Belfast, United Kingdom, xinchenli@qub.ac.uk

in translation than the existing NMT systems since they are fed with much more data than ever, augmented with the state-of-art technology.

However, some NMT systems like DeepL, which is developed by a German company and declared to be the most accurate machine translation ever, still excel in translation. Released in 2017, DeepL is claimed to have surpassed all the other NMT systems in self-conducted blind tests and BLEU scores, including Google Translate, Amazon Translate, Microsoft Translator and Facebook's NMTs (DeepL, no date). A study by Esperança-Rodier and Frankowski (2021) also demonstrates that DeepL has a better performance than Google Translate when translating MWEs (multiword expressions) from French into Polish. With a focus on the Chinese-English language pair, this study compares the translation quality of the two LLMs - ChatGPT 3.5 and ERNIE Bot 3.5, and the NMT system of DeepL to see whether the former have already defeated the latter in the field of translation. It is also going to explore how genres and translation directions may influence these LLMs and NMT systems' translation quality so that hopefully the findings may serve as a reference for users to choose the best translation system and help developers improve their machine's performance in translation.

Previous research on machine translation (MT) quality evaluation has been approached from three primary perspectives. Firstly, the development of models for assessing MT quality which do not focus on any specific system but rather aim to establish a general model applicable to all MT systems. For instance, Liu and Gildea (2005) highlighted the benefits of incorporating syntactic information into evaluation metrics, while Papineni et al. (2002) introduced the BLEU method for assessing translation quality. Furthermore, Wong and Kit (2012) expanded these metrics to include lexical cohesion, enhancing the evaluation at the document level. Lin and Och (2004) introduced the ORANGE method, designed to correlate evaluation scores closely with the quality of translations, and conducted extensive testing on this approach.

The second perspective focuses on the evaluation of specific MT systems to assess their performance from various angles, such as fluency and fidelity. Researchers have compared these targeted systems against other MT systems to benchmark their capabilities. For example, Takakusagi et al. (2021) examined the reliability of DeepL in medical translations from Japanese to English, noting its high accuracy. Additionally, Lyu et al. (2023) explored novel uses for MT using large language models (LLMs), including considerations of privacy concerns. Similarly, Hendy et al. (2023) evaluated the translation performance of GPT from various perspectives, including its comparison with other commercial neural machine translation (NMT) systems and the effectiveness of different prompting strategies. Despite these comprehensive studies, there remains a notable gap in research directly comparing LLMs with traditional NMT systems, particularly in how different genres and translation directions might affect their performance, and in translations between language families such as English and Chinese. This study aims to address these gaps, providing a deeper understanding and a broader evaluation of MT capabilities.

B. Methods

Sampling

This study chose two LLMs - ChatGPT 3.5, ERNIE Bot 3.5, and the NMT system of DeepL to undergo comparison. To test their translation quality, 40 source texts were selected to be fed into these machines, with 20 in Chinese and 20 in English (See Table 1). For each language, there were 10 non-literary excerpts with 5 of them being speech drafts and the rest news reports,

and 10 literary excerpts with half of them being prose and the other half novels. The English or Chinese TTs of all the STs, which were provided by human translators and of high quality, were chosen as frame of reference in the phase of automatic quality evaluation.

Table 1. Basic information of STs

Chinese ST	Genre	Word Count (Characters)	English ST	Genre	Word Count (Words)
黎明前的北京	prose	950	At Sea	prose	345
朋友	prose	1354	Insouciance	prose	347
秋天	prose	1705	On Going a Journey	prose	358
我的父母之乡	prose	916	The Ephemera	prose	340
养花	prose	909	The Windmill	prose	364
边城	novel	490	Eat, Pray, Love	novel	379
三体	novel	964	Holmes	novel	358
三体2	novel	1191	The Hitchhiker's Guide to the Galaxy	novel	357
三体3	novel	1119	The Kite Runner	novel	355
三体4	novel	1589	The Last Battle	novel	364
2016中国的航天	speech	1157	Obama's Inaugural Speech	speech	336
团结协作 开放包容 建设安全稳定、发展繁荣的共同家园	speech	1841	Kennedy's Inaugural Speech	speech	360
发展权：中国的理念、实践与贡献	speech	1675	Biden's Inaugural Speech	speech	367
永远的朋友 真诚的伙伴	speech	1904	Trump's Inaugural Speech	speech	322
在第十一届夏季达沃斯论坛开幕式上的致辞	speech	1709	Johnson's Inaugural Speech	speech	344
各地增植补绿	news	535	From Right-Hand Man to Critical Witness	news	322
暑期档票房创中国影史新纪录	news	482	Global wealth fall	news	336
新时代新征程新伟业	news	375	How Fire Turned Lahaina into a Death Trap	news	298
中国空间站收获阶段性应用成果	news	353	Over 1 in 10 young adults regularly use e-cigarettes	news	320
中欧班列稳定畅通	news	453	Former President Donald J. Trump pleaded not guilty	news	339
Total		21,671	Total		6,911

Procedure

The study was conducted using a structured procedure that involved several steps to evaluate the translation quality of texts processed by large language models (LLMs) and a neural machine translation (NMT) system. Initially, 40 source texts (STs) were translated by the selected machines. The resulting target texts (TTs) were then assessed for quality using the BLEU scoring system, as developed by Papineni et al. (2002). This evaluation took place on Shiyibao.com, a platform that allows users to compare machine-generated translations with human-generated reference texts, awarding a BLEU score based on how closely the machine

translations resemble the human translations. This method was chosen due to BLEU's proven correlation with human judgments over a large test corpus, suggesting its reliability in translation quality assessment.

Subsequently, the BLEU scores obtained were analyzed statistically using SPSS to determine if there were significant differences across various dimensions. The analysis involved two main statistical tests: the T-test and the Kruskal-Wallis Test. The selection of the appropriate test was based on preliminary tests for normality and homogeneity of variances conducted through the Shapiro-Wilk test and variance homogeneity test respectively. If the data were normally distributed with homogeneous variances, a T-test was applied; otherwise, the Kruskal-Wallis Test was used to assess the differences in translation quality among the translations produced by the different systems.

Dimensions of Comparison

The experiment was designed to assess the translation quality across three distinct variables: the type of machine translation systems (machines), the nature of the texts (genres), and the linguistic direction of translation (translation directions). The machines tested included ChatGPT-3.5, ERNIE Bot 3.5, and DeepL. Texts were categorized into literary and non-literary genres, and translations were analyzed between Chinese to English and English to Chinese. The BLEU scores of the target texts (TTs) were compared across three specific dimensions, each isolating different variables to maintain certain constants for controlled comparison.

In the first dimension, the experiment sought to identify potential differences in translation quality among the different machines when translating texts of the same genre in the same translation direction. This analysis aimed to isolate the performance of each machine under uniform conditions of genre and linguistic direction. The second dimension focused on how the translation quality varied when each machine translated texts of different genres but within the same translation direction, providing insights into the impact of genre on machine translation performance. Finally, the third dimension examined the translation quality of each machine working with the same genre but in opposing translation directions, to determine if and how the direction of translation influences the effectiveness of the machine translation systems. This structured approach allowed for a comprehensive understanding of the factors affecting translation quality in machine translation systems.

C. Findings and Discussion

The mean BLEU scores for the TTs are shown in Table 2 below, where the mean for the TTs in each direction (Chinese-English and English-Chinese), of each genre (literary texts and non-literary texts) and of each sub-type under the genre (prose, novel, speech and news) are separately calculated. The findings that are revealed through SPSS analysis in all the three dimensions will be reported respectively in this section.

Table 2. Mean BLEU Scores

	ChatGPT	ERNIE	DeepL
Chinese-English translation	0.65	0.66	0.69
Literary texts	0.66	0.61	0.64
Prose	0.65	0.60	0.62
Novel	0.68	0.63	0.66
Non-literary texts	0.64	0.72	0.74
Speech	0.63	0.78	0.79

	ChatGPT	ERNIE	DeepL
News	0.66	0.66	0.69
English-Chinese translation	0.30	0.31	0.31
Literary texts	0.23	0.22	0.23
Prose	0.21	0.19	0.22
Novel	0.26	0.24	0.24
Non-literary texts	0.37	0.40	0.39
Speech	0.32	0.33	0.36
News	0.42	0.47	0.42

Comparison of Translation Quality in Terms of the Same Genre and Same Direction

Dimension 1 investigates whether there is significant difference of translation quality between different machines when translating the same genre in the same direction. Using T Test and Kruskal-Wallis Test with significance levels both being 0.050, we get the results as are shown in table 3.

Table 3. Results for dimension 1

	ChatGPT-ERNIE	ChatGPT-DeepL	ERNIE-DeepL
Chinese-English Literary	p=0.112 (K-W Test)	p=0.650 (K-W Test)	p=0.699 (T Test)
Chinese-English Non-literary	p=0.316 (T Test)	p=0.597 (K-W Test)	p=0.627 (T Test)
English-Chinese Literary	p=0.463 (T Test)	p=0.863 (T Test)	p=0.570 (T Test)
English-Chinese Non-literary	p=0.558 (T Test)	p=0.677 (T Test)	p=0.786 (T Test)

The results indicate that all the three machines' BLEU scores show no significant difference (p＞0.050) when they translate materials of the same genre in the same direction. This signifies that the recent AI technology, with which ChatGPT 3.5 and ERNIE Bot 3.5 were developed, does not necessarily lead to significant improvement in translating texts of the same genre and in the same direction as compared with DeepL.

Comparison of Translation Quality in Terms of the Same Direction But Different Genres

Dimension 2 explores whether the significant difference exists between the translation quality of each machine when working in the same direction translating texts of different genres. T Test and Kruskal-Wallis Test were conducted to shed light on potential differences, for which the results are presented in Table 4. The significant levels of both tests were also 0.050.

Table 4. Results for dimension 2

	Chinese-English (literary v. non-literary)	English-Chinese (literary v. non-literary)
ChatGPT	p=0.880＞0.050 (K-W Test)	p=0.000＜0.050 (T Test) Literature mean=0.23 Non-literature mean=0.37 Performing better in non-literature
ERNIE	p=0.090＞0.050 (T Test)	p=0.000＜0.050 (T Test) Literature mean=0.22 Non-literature mean=0.40 Performing better in non-literature

	Chinese-English (literary v. non-literary)	English-Chinese (literary v. non-literary)
DeepL	p=0.034＜0.050 (K-W Test) Literature mean=0.636 Non-literature mean=0.742 Performing better in non-literature	p=0.000＜0.050 (T Test) Literature mean=0.23 Non-literature mean=0.39 Performing better in non-literature

In this dimension, the findings demonstrate that the two LLMs show no significant difference in translating literary and non-literary Chinese texts into English (p＞0.050), but DeepL performs better in translating non-literary Chinese texts than literary Chinese texts (p＜0.050). Besides, all the three machines tend to be better at non-literary translation in English-Chinese translation (p＜0.050). To be more specific, the quality of ChatGPT 3.5 does not significantly vary when translating different genres from Chinese into English, but it performs better in non-literary translation when working in the English-Chinese direction. The results for ERNIE-Bot 3.5 are the same as ChatGPT 3.5, while DeepL demonstrates better quality in both translation directions in translating non-literary texts.

Comparison of Translation Quality in Terms of the Same Genre But Different Directions

By conducting T Test and Kruskal-Wallis Test, with both significance levels being 0.050, Dimension 3 compares the translation quality of each machine when they work on the same genre in different directions. Table 5 presents the results of the comparison.

Table 5. Results for dimension 3

	ChatGPT (C-E v. E-C)	ERNIE (C-E v. E-C)	DeepL (C-E v. E-C)
Literary	p=0.000＜0.050 (K-W Test) C-E mean=0.66 E-C mean=0.23 Performing better in C-E	p=0.000＜0.050 (T Test) C-E mean=0.61 E-C mean=0.22 Performing better in C-E	p=0.000＜0.050 (T Test) C-E mean=0.64 E-C mean=0.23 Performing better in C-E
Non-literary	p=0.002＜0.050 (K-W Test) C-E mean=0.64 E-C mean=0.37 Performing better in C-E	p=0.000＜0.050 (T Test) C-E mean=0.72 E-C mean=0.40 Performing better in C-E	p=0.000＜0.050 (T Test) C-E mean=0.74 E-C mean=0.39 Performing better in C-E

These results can prove that all the chosen machines, be it the LLMs or NMT system, have better quality in Chinese-English translation than English-Chinese translation, despite the types of texts on which they are working.

In the first dimension of the study, a comparison was made between different machine translation systems when translating the same genre in the same direction. Surprisingly, all three machines—ChatGPT, ERNIE Bot, and DeepL—demonstrated similar levels of translation quality, challenging the expectation that LLMs, with their advanced AI technology, would outperform traditional NMT systems. This could be due to the traditional NMT systems being specifically trained on large, curated corpora of bilingual text pairs, which enhances their translation specificity, as noted by Forcada (2017). In contrast, LLMs are trained on more generalized data, which might not be specifically aimed at optimizing translation tasks despite their technological sophistication.

In the second dimension, the machines were evaluated on their ability to translate different genres within the same linguistic direction. Here, the results showed a stronger performance in

non-literary translations over literary ones. This is likely because literary texts involve complex emotional nuances and rhetorical elements that are challenging for MT systems to capture. Influential studies, like those by Voigt and Jurafsky (2012), and opinions from Matusov (2019), support this finding by highlighting the inherent difficulties in applying MT to literary translation, with many experts considering it unreliable for such tasks due to its nuanced demands.

The third dimension assessed how well each machine handled translations of the same genre in different directions. Interestingly, all machines performed better in translating from Chinese to English than from English to Chinese, regardless of the genre. This superior performance in one direction might be attributed to better training data available for Chinese-English translation. This finding seems contradictory to the view presented by Chang, Jurafsky, and Manning (2009), who noted that structural and linguistic complexities make Chinese-English translation particularly challenging, typically resulting in lower BLEU scores. However, this apparent contradiction could actually highlight the significant impact of training corpora quality on translation outcomes. It suggests that the machines tested may have had access to a higher quality Chinese-English corpus compared to the English-Chinese corpus, leading to unexpectedly high performance in that direction despite broader challenges in the language pair.

D. Conclusion

This study evaluates the translation quality of two LLMs and one NMT system by comparing the BLEU scores of each TT they produced employing a statistical method. In conclusion, the translation quality of the three machines resembles each other when translating STs of the same genre in the same direction. However, while the two LLMs tend to produce translations of equal quality when working on both literary and non-literary texts in the Chinese to English direction, DeepL performs better in non-literary translation than literary translation in the same direction. In the English to Chinese direction, on the other hand, all the machines produce translations of better quality in non-literary translation than in literary translation. As a whole, all the three machines are more capable of Chinese-English translation than English-Chinese translation. These findings suggest great potential of LLMs working as a machine translator in the future though they currently fail to surpass NMT systems in some aspects. Users may take the results as reference to choose which machine to use, and developers may also find them helpful in improving machine translation quality.

This is just an exploratory study based on a small sampling, but we can still to some extent tell that the current advancement of AI does not necessarily lead to a surge in translation quality. It may therefore not make much sense to switch the NMT we currently use to the LLMs. This finding may serve as a reference for users to choose MT wisely. From another perspective, however, LLMs are still the state-of-the-art AI technology since they are not only competitive in translating all kinds of texts, but also extremely multifunctional to cope with various other tasks, which NMT systems fail to do. They have such huge potential that hopefully they may surpass NMT in the field of translation one day. In addition, these machines have a tendency to perform better in a certain aspect like literature or non-literature, and English-Chinese or Chinese-English direction, and this is probably because the corpora fed into them are different. This finding may provide implications for the developer to optimize their translation products by enhancing the quality of the bilingual corpus to feed the machine, just as we may learn from

the results of Dimension 3 that a different corpus fed may lead to disparity in translation quality in different fields of translation.

References

Chang, P.-C., Jurafsky, D. & Manning, C. D. (2009). Disambiguating "DE" for Chinese-English machine translation. *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pp.215-223.

DeepL. *DeepL Translate: The world's most accurate translator*. Available at: https://www.deepl.com/en/translator

Esperança-Rodier, E. & Frankowski, D. (2021). DeepL vs Google Translate: Who's the best at translating MWEs from French into Polish? A multidisciplinary approach to corpora creation and quality translation of MWEs. *Translating and the Computer 43*. Available at https://hal.science/hal-03779450v1

Forcada, M. L. (2017). Making sense of neural machine translation. *Translation Spaces, 6* (2), 291-309.

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M. & Awadalla, H. H. (2023). How good are GPT models at machine translation? A comprehensive evaluation. *ArXiv.* doi:10.48550/arXiv.2302.09210

Lin, C.-Y. & Och, F. J. (2004). ORANGE: A method for evaluating automatic evaluation metrics for machine translation. *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pp.501-507.

Liu, D. & Gildea, D. (2005). Syntactic features for evaluation of machine translation. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp.25-32.

Lyu, C., Xu, J. & Wang, L. (2023). A paradigm shift: The future of machine translation lies with Large Language Models. *ArXiv.* doi: 10.48550/arXiv.2305.01181

Matusov, E. (2019). The challenges of using neural machine translation for literature. *Proceedings of the Qualities of Literary Machine Translation*, pp.10-19.

Omar, A. & Gomaa, Y. (2020). The machine translation of literature: Implications for translation pedagogy, *International Journal of Emerging Technologies in Learning, 15* (11), 228-235.

Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp.311-318.

Sipayung, K. T., Sianturi, N. M., Arta, I. M. D., Rohayati, Y. & Indah, D. (2021). Comparison of translation techniques by Google Translate and U-dictionary: How differently does both machine translation tools perform in translating? *Elsya: Journal of English Language Studies, 3* (3), 236-245.

Takakusagi, Y., Oike, T., Shirai, K., Sato, H., Kano, K., Shima, S., Tsuchida, K., Mizoguchi, N., Serizawa, I. & Yoshida, D. (2021). Validation of the reliability of machine translation for a medical article from Japanese to English using DeepL translator. *Cureus, 13* (9). doi: 10.7759/cureus.17778

Toral, A., Oliver, A. & Ballestín, P. R. (2020). Machine translation of novels in the age of transformer. *ArXiv.* doi: 10.48550/arXiv.2011.14979

Voigt, R. & Jurafsky, D. (2012). Towards a literary machine translation: The role of referential cohesion. *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pp.18-25.

Wong, B. T. & Kit, C. (2012). Extending machine translation evaluation metrics with lexical cohesion to document level. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp.1060-1068.