ISSN: 2688-8653

Evaluating the Effectiveness of Machine Translation Tools in Translating China-specific Discourse

Liwei Zou

School of Languages and Literature, University of South China, Hengyang 421001, China

Abstract

The accuracy of China-specific discourse translation has emerged as a critical factor in conveying China's voice and shaping its global image with China's comprehensive national strength steadily rising and international exchanges becoming increasingly frequent. This study conducts a comparative analysis between the official English translations of ten representative Chinese sentences from a selected China-specific discourse and the outputs of four mainstream international and domestic machine translation tools: Google Translate, ChatGPT, Wenxin Yiyan (Ernie Bot), and DeepSeek. The objective is to assess the effectiveness of these machine translation tools in the context of China-specific document translation. The evaluation combines automatic metrics—BLEU and TER—to quantify translation quality, and is further supplemented by human judgment to offer a more comprehensive understanding of each system's performance. The results indicate that ChatGPT is the most robust overall, excelling in both machine and human metrics. ERNIE Bot shows strong potential in human-preferred translation of idiomatic and China-specific language, despite poor automatic scores. Finally, this study explores future trends in translation development in the era of artificial intelligence.

Keywords

Machine translation, China-specific discourse, translation quality assessment, translation.

1. Introduction

Translation is far more than a simple process of converting one language into another. From the strategic perspective of enhancing national cultural soft power, it has become a vital link in cultural exchange and knowledge dissemination. In particular, the translation of China-specific texts functions as a crucial channel for governments to engage in external communication, promote national policies and governance philosophies, as well as present the country's social, economic, diplomatic, and cultural conditions to the international community. In recent years, advances in machine translation—especially neural machine translation (NMT)—have significantly improved translation quality and ushered in profound changes in language, culture, and translation studies. However, the application of machine translation to China-specific texts involves more than linguistic transformation; it also demands careful consideration of cultural backgrounds, China-specific contexts, and other complex factors. This necessitates a further investigation into the quality of such translations. Objective and reliable translation quality evaluations can help users make more informed and effective choices regarding machine translation engines and "correct inaccurate machine-generated translations and their potential risks" [1].

Against this backdrop, the present study takes a keynote address—"Hands on the Path Towards Modernization"—as its case study. It conducts a comparative analysis of translations generated by four mainstream machine translation tools: Google Translate and ChatGPT from abroad, and

ISSN: 2688-8653

ERNIE Bot and DeepSeek from China. By employing automated evaluation metrics such as BLEU and TER, and using Python to calculate the corresponding values, the study also incorporates human evaluation to conduct a comprehensive and multi-perspective assessment of translation quality. The goal is to analyze the strengths and limitations of these tools and explore whether they can effectively contribute to telling China's story well. This research aspires to provide valuable insights into the application of machine translation to China-specific texts and to offer a forward-looking perspective on the role of translation in the era of artificial intelligence.

2. Literature Review

A growing number of scholars have begun to examine the opportunities and challenges that generative artificial intelligence poses for language education and academic writing. "AI chatbot and ChatGPT in particular appear to be useful tools in scientific writing, assisting researchers and scientists in organizing material, generating an initial draft and/or in proofreading", however, "several ethical issues arise about using these tools, such as the risk of plagiarism and inaccuracie."[2]

Yuan Yulin[3] provides a detailed account of the principles, algorithms and relevant experiments related to ChatGPT, offering a comprehensive analysis of its potential and limitations in the field of natural language processing, including translation. The machine translation generated by ChatGPT go beyond those of conventional systems, making its performance in translation quality, proofreading capabilities, and sentence optimization a new focal point for research in both machine translation and translation studies[4]. The existing literature on ChatGPT and translation primarily centers on two key areas. The first focuses on translation quality assessment based on automatic evaluation metrics such as BLEU and TER. For instance, Yang Fengchang[5] used ChatGPT to translate Vietnamese legal texts and, through comparison with outputs from other machine translation tools and human translations, identified its strengths and weaknesses, further reflecting on the implications of ChatGPT for translators. Wen Xu and Tian Yaling[6] compares the translation generated by ChatGPT of The Report to the 20th National Congress with that of three mainstream machine translation tools—Google Translate, Youdao Translate, and DeepL Translate. They found out that the results indicate that ChatGPT exhibits certain advantages compared with the other three translation tools, although it still has obvious limitations. The second area concerns the advantages and challenges that AI chatbot brings to language and translation. Wang Lifei and Li Zhao[7] highlight ChatGPT's role as an AIGC (Artificial Intelligence Generated Content) tool in accelerating the emergence of an era of intelligent translation and language services. They propose the introduction of translation industry-relevant courses to facilitate a shift from traditional translation instruction to language service training and to cultivate translators' digital humanities literacy. Zhu Guanghui and Wang Xiwen[8] argue that ChatGPT's AI-driven operational model can provide support for simultaneous interpreting and translation in multilingual education settings.

As one of the earliest and most representative tools in the field of generative artificial intelligence, ChatGPT has been at the forefront of academic research since its launch. Existing studies have primarily focused on its disruptive impact on the translation industry or evaluated its translation quality using multidimensional metrics. However, with the rapid advancement of artificial intelligence technologies in China, the development and application of domestic large language models have made significant breakthroughs. Domestic intelligent engines, exemplified by DeepSeek, have demonstrated continuous improvements in performance across natural language processing tasks. Despite this progress, comparative research on the performance of domestic AI engines and internationally mainstream models in specific translation domains remains limited. In particular, within the field of China-specific text

ISSN: 2688-8653

translation—which demands not only fidelity but also heightened China-specific awareness—there is a notable lack of systematic investigation into the effectiveness of domestic AI tools. This represents a significant research gap in current scholarship.

3. Research Design

This study aims to address the following three research questions: First, what is the quality of translations generated by machine translation tools? Second, do these tools have the potential to replace human translators in specific types of translation tasks? Third, how might translation practices evolve in the era of artificial intelligence?

The source text selected for this study is President Xi's keynote address entitled "*Hands on the Path Towards Modernization*". The speech is rich in discourse with distinct Chinese characteristics and represents a unique form of linguistic expression rooted in China-specific and cultural traditions. A qualitative analysis of the speech's discourse reveals three main features. First, the language demonstrates a high degree of structural regularity. At the lexical level, the speech frequently employs four-character phrases, which convey meaning precisely and concisely. At the syntactic level, parallel and coordinate sentence structures are commonly used. Second, the speech embodies profound China-specific thought. As Zhu Yihua[9] suggests, publicity translation should pay attention to national interests, China-specific viewpoints, and ideological considerations, highlighting its inherently China-specific nature. The speech, themed "The Path to Modernization: The Responsibility of China-specific Parties," offers a Chinese solution to shared global challenges related to modernization. Third, the text is rich in China-specific metaphors. Such metaphors are deeply embedded in traditional Chinese culture and values, serving not only to evoke emotion and shape ideology but also as vital tools for telling China's story and amplifying its voice internationally. Based on these characteristics, this study selects ten representative sentences from the speech for analysis, comprising 425 Chinese characters and 270 words in the official English translation.

Table 1. Ten Example Source Sentences

例一：当今世界，多重挑战和危机交织叠加，世界经济复苏艰难，发展鸿沟不断拉大，生态环境持续恶化，冷战思维阴魂不散，人类社会现代化进程又一次来到历史的十字路口。
例二：我们要坚守人民至上理念，突出现代化方向的人民性。
例三：现代化不是少数国家的“专利品”，也不是非此即彼的“单选题”，不能搞简单的千篇一律、“复制粘贴”。
例四：人类是一个一荣俱荣、一损俱损的命运共同体。
例五：吹灭别人的灯，并不会让自己更加光明；阻挡别人的路，也不会让自己行得更远。
例六：要坚持共享机遇、共创未来，共同做大人类社会现代化的“蛋糕”，努力让现代化成果更多更公平惠及各国人民.
例七：现代化不会从天上掉下来，而是要通过发扬历史主动精神干出来。
例八：一花独放不是春，百花齐放春满园。
例九：我们要共同倡导尊重世界文明多样性，坚持文明平等、互鉴、对话、包容，以文明交流超越文明隔阂、文明互鉴超越文明冲突、文明包容超越文明优越。
例十：中国愿同各方一道努力，让各具特色的现代化事业汇聚成推动世界繁荣进步的时代洪流，在历史长河中滚滚向前、永续发展！

ISSN: 2688-8653

4. Research Methods

At present, the most widely used automatic evaluation metric in the machine translation industry is BLEU (Bilingual Evaluation Understudy) [10]. This metric assesses the quality of machine-generated translations by comparing them with human reference translations. BLEU primarily relies on n-gram matching, which measures the frequency with which n-grams in the machine translation appear in the reference translations. Scores typically range from 0 to 1, with higher values indicating better translation quality. BLEU is favored for its simplicity and computational efficiency. Despite its limitations, it remains one of the most commonly adopted metrics for evaluating machine translation performance. Another widely used metric is TER (Translation Edit Rate), which differs from BLEU in that it is based on edit distance. In essence, TER quantifies the minimum effort of editing required to make the machine translation identical to the reference translation. Like BLEU, TER scores range from 0 to 1, but higher scores indicate lower translation quality. Due to its sensitivity to fine-grained differences between translations, TER serves as a valuable complementary metric.

By combining these two distinct evaluation approaches, a more comprehensive and precise assessment of machine translation quality can be achieved. In addition, the present study employs human evaluation to supplement the automatic metrics. Ten translated sentences generated by four different translation tools were anonymized and distributed to two professional translators. Without knowledge of the translation sources, the evaluators rated the quality of each translation. The overall scores for the ten sentences were then calculated and analyzed, providing a multi-faceted and multi-perspective evaluation of the performance of machine translation models.

5. Results and Discussion

This study uses the prompt “Please provide the English translation for these sentences” to elicit translations from four mainstream machine translation tools, which are then compared across. For each translation, BLEU and TER scores were calculated to evaluate translation quality. The BLEU and TER results for the selected ten examples, computed using Python, are presented in Figures 1 and 2, respectively.

Figure 1. BLEU scores of the four machine translation tools

Tool	Example 1	Example 2	Example 3	Example 4	Example 5
Google	0.5317	0.0460	0.2781	0.3016	0.0320
ChatGPT	0.2977	0.1703	0.2243	0.2110	0.0150
ERNIE Bot	0.1396	0.0486	0.2220	0.0187	0.0144
DeepSeek	0.0738	0.0617	0.0464	0.0073	0.0564
Tool	Example 6	Example 7	Example 8	Example 9	Example 10
Google	0.0603	0.0821	0.4925	0.2845	0.2718
ChatGPT	0.0732	0.1275	0.6268	0.5084	0.2149
ERNIE Bot	0.0270	0.0667	0.4779	0.2415	0.1201
DeepSeek	0.0553	0.0123	0.3822	0.3247	0.1140

ISSN: 2688-8653

Figure 2. TER scores of the four machine translation tools

Tool	Example 1	Example 2	Example 3	Example 4	Example 5
Google	0.3529	0.9167	0.5610	0.5625	0.9286
ChatGPT	0.4706	0.9167	0.6098	0.4375	0.9286
ERNIE Bot	0.6667	10000	0.6585	0.8125	0.9643
DeepSeek	0.7059	0.7500	0.7561	0.8125	0.8929
Tool	Example 6	Example 7	Example 8	Example 9	Example 10
Google	0.9375	0.7273	0.4286	0.5319	0.6842
ChatGPT	0.7188	0.6818	0.2381	0.3191	0.5263
ERNIE Bot	0.7812	0.7273	0.3333	0.6383	0.7105
DeepSeek	0.7500	0.8636	0.3810	0.5532	0.6316

Figure 3. Average BLEU Scores Across All 10 Examples

Tool	Average BLEU Score
ChatGPT	0.2970
Google	0.2381
ERNIE Bot	0.1388
DeepSeek	0.1138

Figure 4. Average TER Scores Across All 10 Examples

Tool	Average BLEU Score
ChatGPT	0.5486
Google	0.6365
DeepSeek	0.6620
ERNIE Bot	0.7323*

*ERNIE Bot has one anomalous TER value ("10000") in example 2, likely a placeholder for failure. Excluding it, the adjusted average is 0.6379.

The BLEU and TER scores across ten translation examples reveal significant differences in the performance of the four machine translation tools—Google, ChatGPT, ERNIE Bot, and DeepSeek. In terms of BLEU scores, which reflect lexical and phrasal overlap with reference translations, ChatGPT outperforms all others in BLEU, indicating better n-gram overlap with human reference. Google also stands out with nearly identical average scores, indicating that both systems are generally capable of producing fluent and faithful outputs at the surface level. ChatGPT achieves particularly high BLEU scores in Examples 8 and 9, suggesting strong performance on those inputs. In contrast, ERNIE Bot and DeepSeek perform considerably worse, with DeepSeek consistently generating the lowest BLEU scores across most examples. This indicates a weaker ability to reproduce n-gram patterns consistent with human translations. The TER scores, which measure the amount of editing needed to align machine translations with human references, offer a complementary perspective. ChatGPT again leads with the lowest TER, indicating its outputs require fewer edits to meet human standards. Google ranks second in this respect, with a slightly higher average TER, followed by DeepSeek and ERNIE Bot. Interestingly, while Google performs well in BLEU, its higher TER in some examples—such as Examples 2, 5 and 6—suggests that its translations, while lexically similar, may deviate structurally or grammatically. TER results also highlight certain problematic

ISSN: 2688-8653

examples (e.g., Example 2), where ERNIE Bot scores a full 1.0, indicating a complete mismatch with the reference.

Overall, the combined analysis of BLEU and TER indicates that ChatGPT provides the most balanced translation performance—both in terms of textual similarity and minimal post-editing effort. Google follows closely, showing strength in BLEU but less consistency in TER. ERNIE Bot and DeepSeek appear to struggle in both metrics, producing translations that are not only less similar to references but also require more extensive human correction. These findings underscore the importance of using both surface-level and structure-sensitive metrics in evaluating machine translation quality, as relying on BLEU or TER alone may obscure important aspects of translation adequacy and fluency.

While BLEU and TER scores provide a quantitative framework for assessing translation quality, human evaluation offers a more nuanced understanding by capturing elements such as accuracy, stylistic appropriateness, and cultural adaptation—factors often beyond the reach of automated metrics. To ensure a more holistic assessment, this study combines both approaches. Two senior professionals specializing in translation were invited to conduct the human evaluation, with their assessments presented in Figures 5 and 6, respectively. Each translation was scored manually by human raters out of 100.

Figure 5. Scores from the First Human Evaluation

Tool	Google	ChatGPT	ERNIE Bot	DeepSeek
Example 1	88	90	94	90
Example 2	88	90	94	94
Example 3	84	92	94	84
Example 4	90	94	90	90
Example 5	80	94	92	80
Example 6	84	94	94	92
Example 7	82	84	94	92
Example 8	84	88	92	92
Example 9	82	92	90	92
Example 10	84	94	92	92

Figure 6. Scores from the Second Human Evaluation

Tool	Google	ChatGPT	ERNIE Bot	DeepSeek
Example 1	80	87	90	85
Example 2	89	95	85	85
Example 3	55	90	85	75
Example 4	80	85	75	70
Example 5	80	78	65	55
Example 6	80	75	65	85
Example 7	83	90	85	80
Example 8	75	85	90	80
Example 9	80	78	90	85
Example 10	70	75	78	76

ISSN: 2688-8653

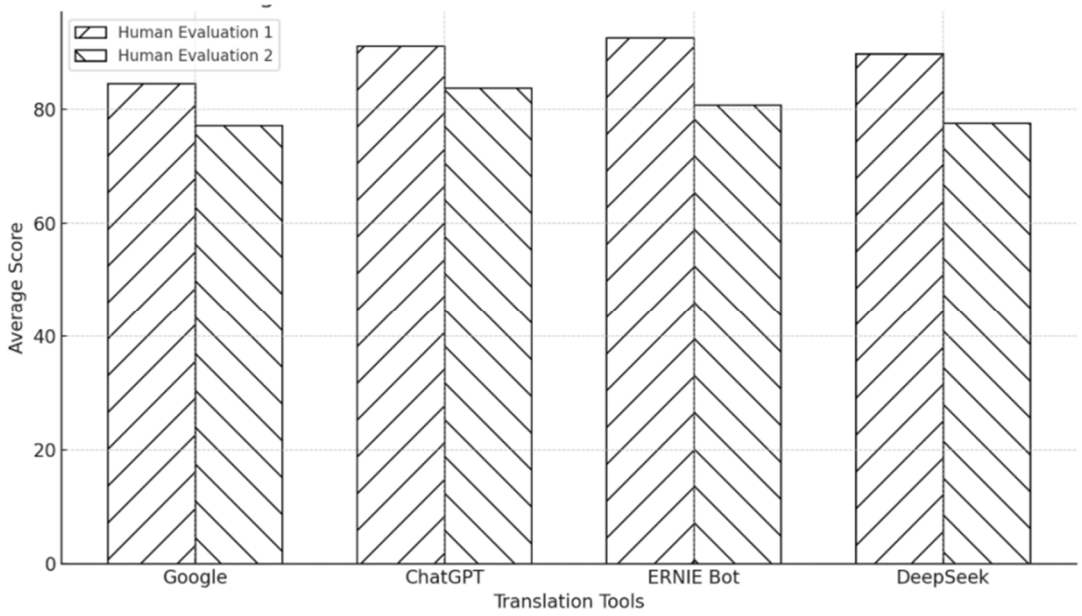

Figure 7. Average Scores of Translation Tools in Two Human Evaluations

The statistical results from the two rounds of human evaluation reveal notable differences in translation performance among the four tools. Overall, ERNIE Bot achieved the highest average score in the first evaluation (92.6), while ChatGPT demonstrated strong and consistent performance across both evaluations (91.2 and 83.8). In contrast, Google and DeepSeek received relatively lower scores, particularly in the second evaluation, with averages of 77.2 and 77.6 respectively. The second reviewer assigned generally lower scores to all tools, suggesting stricter evaluation standards or differing assessment criteria. The consistent lead of ChatGPT over Google by a margin of more than six points in both evaluations indicates its stable quality advantage. While ERNIE Bot showed top-tier performance in one evaluation, ChatGPT's overall consistency positions it as the most reliable tool across evaluators. These findings highlight the importance of combining human and automatic evaluation methods to capture both quantitative accuracy and more nuanced aspects of translation quality. Interestingly, ERNIE Bot received the highest average human scores, even though it had lower automatic metrics. This suggests that it may perform better on stylistic or culturally nuanced content, which automatic metrics don't fully capture.

Overall, ChatGPT is the most balanced performer, scoring consistently well in both BLEU and TER, and very well in human evaluation. Google did moderately in BLEU and TER, but ranked lowest in human judgments. ERNIE Bot underperforms in automated metrics, but was preferred by human evaluators—potentially due to better handling of Chinese cultural or idiomatic expressions. DeepSeek trails overall but performed better than ERNIE Bot in BLEU/TER, showing strength in literal or structural translation.

6. Conclusion

Following a comprehensive evaluation of four major machine translation tools, this study finds that, from a quantitative perspective, ChatGPT demonstrates superior performance in Chinese-to-English translation of China-specific texts. Human evaluation results also indicate that ERNIE Bot and ChatGPT outperforms other mainstream translation tools currently available on the market. These findings suggest that, as the GenAI model continues to be refined, it holds significant potential for broader application in the field of machine translation. In the context of post-editing, ChatGPT may serve as an effective pre-processing tool by generating higher-quality initial translations, thereby improving overall translation efficiency and quality.

ISSN: 2688-8653

Nonetheless, for many specialized and high-stakes translation tasks, a hybrid approach combining human translators and machine translation remains the optimal solution. In the era of artificial intelligence, the field of translation should prioritize the following areas for development: (1) Technological integration, leveraging modern technologies to enhance translation efficiency and accuracy; (2) Ethical inquiry, addressing the ethical challenges associated with machine translation and automation to ensure legal and morally responsible practices; (3) Cross-cultural research to deepen the understanding of cultural differences and linguistic diversity to better meet translation needs in multilingual and multicultural contexts; (4) Human-machine collaboration in exploring synergistic workflows between human translators and AI tools to enhance translation quality, such as integrating machine translation with human post-editing; (5) Educational innovation to update translation education and training to equip students and professionals with the skills to effectively use modern translation technologies; and (6) Technological research for investigating both the potential and limitations of current translation technologies to drive future advancements.

Acknowledgments

This work was financially supported by the fund of the key scientific research project of the Hunan Provincial Education Department “Construction and Application of Human-Computer Interactive Collaborative Translation Model in the Era of ChatGPT” (No. 23A0321)

References

[1] Zhang, X. J., & Shao, L. The basic principles for building trustworthy machine translation systems: An engineering ethics perspective [J]. Foreign Languages and Literature, 2021, (1), p.1–8.

[2] Salvagno, M., Taccone, F. S., & Gerli, A. G. Can artificial intelligence help for scientific writing? [J]. Critical Care, 2023, 27(1), p.75.

[3] Yuan, Y. L. Beyond chatbots: The success of ChatGPT and its implications for linguistics [J]. Contemporary Linguistics, 2023, (5), p.633–652.

[4] Geng, F., & Hu, J. New directions in AI-assisted post-editing: A case study using ChatGPT [J]. China Foreign Languages, 2023, (3), p.41–47.

[5] Yang, F. C. Reflections and implications of ChatGPT for interpreters: A case study of Vietnamese legal translation [J]. Chinese Science & Technology Translators Journal, 2023, (3), p.27–30, 4.

[6] Wen, X., & Tian, Y. L. A study on the effectiveness of ChatGPT in translating Chinese characteristic discourse [J]. Shanghai Journal of Translators, 2024, (2), p.27–34.

[7] Wang, L. F., & Li, Z. ChatGPT accelerates the transformation of translation and foreign language education [EB/OL]. https://tech.china.com/article/20230222/022023_1226999.html. (2023-02-22) .

[8] Zhu, G. H., & Wang, X. W. The operational model, key technologies, and future prospects of ChatGPT [J]. Journal of Xinjiang Normal University (Philosophy and Social Sciences Edition), 2023, (4), p.113–122.

[9] Zhu, Y. H. A China-specific analysis of external publicity translation and its translation strategies [M]. Suzhou: Suzhou University Press, 2017, p.1.

[10] Wang, J. S., Zhuang, C. Q., & Wei, Y. P. Machine translation quality assessment: Methods, applications, and outlook [J]. Foreign Languages and Literature, 2024, (5), p.135–144.