ISSN 2414-4746

MODERN VECTORS OF SCIENCE AND EDUCATION DEVELOPMENT IN CHINA AND UKRAINE

中国与乌克兰科学及教育前沿研究

Harbin Engineering University
State institution "South Ukrainian National Pedagogical University named after K. D. Ushynsky"
Educational and Cultural Center "Confucius Institute"

Odesa, Ukraine

Harbin, the People's Republic of China

MODERN VECTORS OF SCIENCE AND EDUCATION DEVELOPMENT IN CHINA AND UKRAINE

中国与乌克兰科学及教育前沿研究

2024
ISSUE № 10

ISSN 2414-4746

MODERN VECTORS OF SCIENCE AND EDUCATION DEVELOPMENT IN CHINA AND UKRAINE

中国与乌克兰科学及教育前沿研究

2024
ISSUE № 10

ISSN 2414-4746

MODERN VECTORS OF SCIENCE AND EDUCATION DEVELOPMENT IN CHINA AND UKRAINE

中国与乌克兰科学及教育前沿研究

The State institution "South Ukrainian National Pedagogical University named after K. D. Ushynsky"

Harbin Engineering University

2024
ISSUE № 10

Odesa, Ukraine

Harbin, the People's Republic of China

MODERN VECTORS OF SCIENCE AND EDUCATION DEVELOPMENT IN CHINA AND UKRAINE **ISSN 2414-4746**

This international journal, as a periodical, includes scientific articles of Ukrainian and Chinese scholars on the problems of Sinology, Cross-cultural Communication, Pedagogy and Psychology: contemporary review. Odesa, Ukraine.

Issue № 10

South Ukrainian National Pedagogical University named after K. D. Ushynsky
Odesa, Ukraine, 2024
Harbin Engineering University
Harbin, the People's Republic of China, 2024

Modern vectors of science and education development in China and Ukraine (中国与乌克兰科学及教育前沿研究): International annual journal. – Odesa: South Ukrainian National Pedagogical University named after K. D. Ushynsky, Harbin: Harbin Engineering University, 2024. – Issue 10. – 390 p.

The ninth issue of the materials represented by the Ukrainian and Chinese scholars are dedicated to the relevent issues of General and Contrastive Linguistics within the Chinese, English, Ukrainian, Turkish and Korean languages; linguodidactic problems of teaching native and foreign languages within polycultural educational space; peculiarities of cross-cultural communication in geopolitical space alongside education-related aspects regarding profession-oriented training of future specialists under conditions of multicultural environment and military actions in Ukraine; post-COVID-19 pandemic chalanges.

The given articles may be of use to researchers, graduates, postgraduates and practising teachers who are interested in various aspects of Sinology, Cross-cultural Communication, Linguistics, Pedagogy and Psychology.

ISSN 2414-4746

Recommended for press
by the Academic Council
(Minute #15 dated 25 April 2024),
South Ukrainian National Pedagogical
University named after K. D. Ushynsky,
Harbin Engineering University

South Ukrainian National Pedagogical University named after K. D. Ushynsky,
Odesa, Ukraine
Harbin Engineering University
Harbin, the People's Republic of China

DOI: 10.24195/2414-4746-2024-10-32
UDC: 81'25:655.535.5-029.32(045)
Ч57

Zhang Xueqing
Graduate Student，School of Foreign Languages,
Harbin Engineering University, Harbin, China
Liang Hong
Professor
School of Foreign Languages,
Harbin Engineering University, Harbin, China

CROSS-MODEL COMPARISON: THE EFFECTIVENESS OF LARGE LANGUAGE MODELS IN TRANSLATING POLITICAL TEXTS

Abstract：The swift evolution of Large Language Model (LLM) technologies has underscored their expansive applicability across a broad spectrum of disciplines, notably within the realms of natural language processing and machine translation. Thus, to comprehensively evaluate the efficacy of machine translation applications in translating political texts under different technological and algorithmic contexts, a curated test dataset comprising 200 typical sentences pertinent to political contexts was developed. leveraging the unique linguistic structural nuances of political texts, four automatic evaluation metrics-BLEU, chrF++， TER, and METEOR-were employed to facilitate both quantitative and qualitative analyses of translations rendered by LLM ChatGPT (4. 0) and ERNIE Bot (4. 0), alongside two leading translation engines: Google Translate and DeepL Translate. The findings not only provide empirical support for understanding the application efficacy of machine translation systems in the field of political translation but also offer insights into algorithm optimization and translation accuracy improvement for developers of machine translation technologies. Furthermore, they provide practical guidance for professionals in selecting appropriate translation systems, which can facilitate

intercultural communication, and provide more comprehensive support for building China's international image.

Keywords: *large language models; automatic evaluation metrics; political texts translation.*

1. Introduction

The field of artificial intelligence is currently experiencing a significant paradigm shift, primarily driven by the rapid advancements in large language models (LLMs) such as OpenAI's GPT series, Google's BERT [13], and similar technologies. These models have shown exceptional capability in various tasks including text generation, language understanding, and machine translation, setting new benchmarks for technological innovation. The emergence of ChatGPT, in particular, has captured the attention of scholars for its potential to revolutionize language teaching and academic writing. Unlike traditional machine translation services, ChatGPT offers enhanced translation quality, proofreading ability, and sentence optimization, marking a pivotal moment for machine translation and translation studies [10]. Jiao et al. (2023) critically assessed ChatGPT's capabilities against commercial translation products like Google Translate, focusing on aspects such as translation prompts, multilingual translation, and robustness [4]. Their findings highlight the superior machine translation effects triggered by specific instructions, underscoring the nuanced capabilities of ChatGPT. Similarly, Liu Shijie (2024) delved into the effectiveness of machine translation in maritime translation, employing metrics such as BLEU, chrF++, and BERTScore for a comprehensive evaluation [6]. Further, a comparative study by Y Sahari et al. on the preferences between ChatGPT and Google Translate among translation teachers and students reveals a divided inclination, with students favoring ChatGPT and teachers opting for Google Translate [10, p. 52]. In addition, Cao, S., & Zhong, L. (2023) have discovered that ChatGPT-based feedback excels in enhancing lexical capability and referential cohesion, while traditional feedback from teachers and students better addresses the development of syntax-related skills [2].

LLM like ChatGPT, Copilot, and ERNIE Bot are great at translating general texts, but how well they translate political texts with cultural and technical language is not

well known. Accurate political text translation is essential because mistakes could cause diplomatic issues. This study compared these models with traditional tools like Google Translate and DeepL Translate using a set of 200 Chinese-to-English political texts. We assessed their translation quality using both automated metrics (BLEU, CHRF++, TER, METEOR) and manual reviews. Our goal is to evaluate how these models handle the unique challenges of political texts in terms of accuracy, fluency, and understanding. This research will improve political text translation with AI, aiming to better build China's image globally.

2 Research Methods

Considering that different "prompts" may lead to different results when translating with AI assistants based on large-scale language models, we adopted a uniform prompt "Please provide [TGT] translations for these sentences:" in order to ensure that translations are provided by state-of-the-art large-scale language models, including ChatGPT (4.0), Copilot, and ERNIE Bot (4.0). It then benchmarks these translations against those produced by conventional machine translation software, namely Google Translate and DeepL Translate, to evaluate their quality. The BLEU, CHRF++, TER, and METEOR scores were calculated for each sentence. Then these scores were averaged for each evaluation metric across all sentences to derive an average score per tool, thus offering a holistic view of each model's performance on the dataset. Upon gathering the scores for BLEU, CHRF++, TER, and METEOR, we undertook a comprehensive assessment of translation quality, integrating these findings with manual evaluations to present a nuanced analysis of translation efficacy.

2.1 Dataset Construction

To constructe a test set for translating political texts from English to Chinese (comprising 200 example sentences), strict standards were applied to select test sentences, carefully chosen from the most representative political texts between 2020 and 2024 to ensure the test set covers key concepts and main scenarios in the field of political texts as comprehensively as possible. The content specifically includes the "*REPORT ON THE WORK OF THE GOVERNMENT 2024*", "*The Belt and Road Initiative: A Key Pillar of the Global Community of Shared Future*", "*Address at the*

Closing Ceremony of the BRICS Business Forum 2023", "Foreign Minister Wang Yi answered questions from Chinese and foreign media".These texts were chosen because their content not only covers a wide range of political issues but also contains complex language features, such as technical terms, political metaphors, and culturally specific expressions. Reference translations are derived from officially published English versions to ensure the standardization and authority of the translations.

2.2 Translation Tools and Model Selection

We selected three mainstream large language models for comparison, including ChatGPT (4.0), Copilot, and ERNIE Bot (4.0), as well as two major translation engines, Google Translate and DeepL Translate. These tools and models were chosen for their widespread application and advancement within the field of machine translation, representing the current pinnacle of artificial intelligence translation technology. All machine translation outputs referenced in this article were generated before March 10, 2024.

2.3 Automatic evaluation indicators

The BLEU metric, proposed by Papineni et al. in 2002, measures machine translation accuracy by comparing the machine's text to reference translations using n-gram overlaps [7, p,314]. Despite its simplicity and being a standard evaluation tool, it has faced criticism for not fully capturing the semantic accuracy and fluency of translations. BLEU scores, which range from 0 (no match) to 1 (perfect match), are often presented as percentages in industry evaluations for easier comparison.

chrF++, introduced by Popovic in 2017, improves upon the chrF metric by focusing on character-level rather than word-level analysis [8, p.312]. This makes it more suitable for languages with significant structural differences or unique expressions, as it can detect finer linguistic nuances like spelling and morphological changes.

TER, proposed by Snover et al. in 2006, calculates the minimum edits needed to match a machine-generated text to a reference translation, providing a ratio that reflects the translation's accuracy [12]. While offering direct insight into the translation's fidelity, TER may not fully account for semantic nuances or accommodate multiple

correct translations.

METEOR, introduced by Banerjee and Lavie in 2005, seeks to address the limitations of other metrics by evaluating translations based on semantic factors [1, p.66], including synonymy and word order. Scoring from 0 to 1, METEOR aims for a more nuanced assessment of translation quality, emphasizing the importance of both accuracy and fluency.

These four metrics offer a comprehensive evaluation framework, allowing for a detailed analysis of machine translation performance, especially in translating complex texts like political documents.

2.4 Human Evaluation

In addition to automatic evaluation, we also employed human evaluation to gain a more comprehensive and in-depth feedback on translation quality. Two professional translation reviewers were invited to conduct blind reviews of the machine translation outputs, coming from diverse backgrounds with extensive translation experience and a profound understanding of political texts. The review criteria included the translation's accuracy, fluency, and the ability to accurately convey the original text's intent and emotional nuances.

3 Research Results

Calculated using Python, the performance of each translation system on the BLEU, CHRF++, TER, and METEOR metrics is shown in Table 1:

Table 1: Scores of each translation system on BLEU, chrF++, TER, and METEOR metrics (rounded to 3 decimal places).

System	BLEU	chrF+ +	TER	METEOR
GPT	40.745	62.891	40.837	67.531
ERNIE Bot	46.265	62.365	54.554	60.373
Copilot	40.496	60.727	41.398	65.348
Google	41.130	57.025	48.775	58.902
DeepL	32.700	43.624	65.345	56.217

GPT excels in chrF++ (62.891) and METEOR (67.531), indicating strong character-level quality and excellent semantic accuracy and fluency.

ERNIE Bot leads in BLEU (46.265), showing top lexical accuracy, but trails in METEOR, suggesting room for improvement in semantic precision and fluency.

Copilot demonstrates a good balance with the lowest TER score, indicating its translations are most akin to reference texts, and a strong METEOR score, reflecting good semantic quality.

Google and DeepL show varied performances across metrics. Google's results suggest moderate performance across the board with room for improvement in semantic accuracy, while DeepL's lowest scores in BLEU and chrF++ and highest in TER indicate it significantly diverges from reference translations in both lexical and character-level accuracy, although it somewhat compensates with a decent METEOR score.

Compared to these four automatic evaluation metrics, manual review can more accurately capture the translation's accuracy, language style, and cultural adaptability, among other qualitative aspects. Both approaches should be used in conjunction to obtain a comprehensive evaluation of translation quality. For this experiment, two senior professors in translation were invited for manual review, with their ratings presented in Table 2.

Table 2: Manual review scores for each example sentence

System	Reviewer 1	Reviewer 2
GPT	78.85	79.71
ERNIE Bot	70.10	72.84
Copilot	70.55	70.32
Google	55.22	51.10
DeepL	55.00	50.91

The analysis of the evaluative scores assigned by the first reviewer indicated that the translations generated by ChatGPT received an average rating of 78.85, outperforming ERNIE Bot (70.10), Copilot (70.55), Google Translate (55.00), and DeepL Translate (55.22) respectively. Consistently, the second reviewer's assessments corroborated this ranking, with ChatGPT achieving an elevated average score of 79.71, thereby exceeding the performance of ERNIE Bot (72.84), Copilot (70.32), Google Translate (50.91), and DeepL Translate (51.10). Despite observable variances in

individual scores, the hierarchical ordering of the translation tools' efficacy remained constant across both reviewers' evaluations, unequivocally establishing a preferential endorsement of ChatGPT's translation capabilities. This consensus underscored ChatGPT's superior proficiency in executing the designated translation task, positioning ERNIE Bot as the subsequent preference, and relegating DeepL Translate to the lowest echelon of performance among the evaluated tools.

4 Discussion

The analysis of both automated and manual evaluations indicates the unique strengths of large language models and translation engines in translating political texts. Notably, ChatGPT stands out with superior performance in CHR++ and METEOR evaluations, highlighting its exceptional translation quality and profound comprehension. While ERNIE Bot's dominance in BLEU scores reflects its notable lexical matching and accuracy. The lower TER scores of DeepL Translate and ERNIE Bot indicate their efficiency in producing outputs closer to reference translations with minimal edits. However, DeepL Translate exhibits weaker performance, particularly in METEOR scores, suggesting potential shortcomings in semantic understanding and fluency.

To further investigate the specific challenges of translating political texts and how these models compare with traditional translation engines, we specifically analyzed sentences where ChatGPT's manual and METEOR scores surpassed those of its counterparts. For instance, in translating the culturally nuanced political term "四风" into English, ChatGPT opted for an expression comprehensible to international readers, accurately conveying the concept of "formalism, bureaucratism, hedonism, and extravagance." This sensitivity and adaptability to cultural differences are hallmarks of high-quality translation, a feat not as well accomplished by other systems. ERNIE Bot's translation of "四风" as "four forms of bad conduct," although relatively accurate, lacks precision. Copilot, Google Translate, and DeepL Translate merely rendered it as "four winds," failing to capture the term's original meaning. ChatGPT's translation is not only faithful to the source but also fluent and natural, reflecting the model's training on extensive corpora, which enables translations that resonate with the target language's

idiomatic expressions.

These observations suggest that although mainstream translation tools offer expedient and accessible services, specific large language models may deliver superior quality in translating political texts that demand high accuracy and comprehension. The translation of "三农" by LLM-based AI systems as "agriculture, rural areas, and farmers," compared to the simplistic "three rural areas" by Google Translate and DeepL Translate, underscores this point. Yet, it is critical to recognize the discrepancy with the official translation, "agriculture, rural areas, and rural residents." The term "rural residents," referring to Chinese citizens registered in rural locales, and the misalignment with the Oxford Dictionary's definition of "farmer," reveal the nuanced understanding still required from LLM-based translations. This necessitates the intervention of human translators for refinement and verification, highlighting the blend of human expertise and AI capabilities in achieving translation excellence.

The findings of this study suggest that in the era of machine translation, the role of human translators is becoming increasingly important, and rather than being marginalized, it is becoming more and more important. ChatGPT has not yet caused a fundamental change to traditional human translators or to current machine translation tools. Especially when dealing with politically sensitive texts, the role of human translation is still critical. As the ChatGPT model continues to be optimized, it will have wide application potential in the field of machine translation, but in many professional and critical translation tasks, the combined approach of human translators and machine translation remains the best choice. Human translators have irreplaceable advantages in interpreting complex contexts, navigating cultural differences, and ensuring the accuracy and consistency of translations. The role of the translator is evolving into that of a quality assurer, responsible for ensuring the accuracy, fluency and cultural appropriateness of the translation. They also act as cultural mediators, preserving and conveying the cultural connotations and emotional color of the original text.

Therefore, in the era of artificial intelligence, especially in the face of natural language processing technologies such as ChatGPT, translators should strengthen the

construction and development of the following aspects: First of all, translators urgently need to establish an effective model of human-machine cooperation with the machine translation system, taking advantage of the efficiency of the machine translation while also utilizing their professional judgment. This paradigm shift requires going beyond traditional roles and expanding the skills, methods, and mediums of translation [6]. Secondly, Fu Jingmin (2023) proposed that Chinese translation education in the new era should be underpinned by the rich essence of traditional Chinese translation culture, and we need to strengthen the Chinese element in translation education, strengthen China's position in the international discourse system through translation, and help disseminate Chinese culture and Chinese voices [4, p.13]. It is also important to focus on the study of translation ethics, especially the ethical challenges associated with machine translation and automated tools, in order to ensure legal and ethical translation practices.

REFERENCES

1. Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization (pp. 65-72).

2. Cao, S., & Zhong, L. (2023). Exploring the effectiveness of ChatGPT-based feedback compared with teacher feedback and self-feedback: Evidence from Chinese to English translation. arXiv preprint arXiv:2309.01645.

3. 傅敬民.(2023).翻译作为独立学科的新时代中国翻译教育.外语电化教学(01),11-13.

4. Jiao, W., Wang, W., Huang, J., et al. (2023). Is ChatGPT a good translator? Yes with GPT-4 as the engine. arXiv preprint arXiv:2301.08745.

5. Lee, T. (2023). Artificial intelligence and posthumanist translation: ChatGPT versus the translator. Applied Linguistics Review. https://doi.org/10.1515/applirev-2023-0122

6. 刘世界.(2024).涉海翻译中的机器翻译应用效能：基于 BLEU、chrF++和 BERTScore 指标的综合评估.中国海洋大学学报(社会科学版)(02),21-

31.doi:10.16497/j.cnki.1672-335X.202402003.
7. Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318).
8. Popović, M. (2017, September). chrF++: words helping character n-grams. In Proceedings of the second conference on machine translation (pp. 612-618).
9. 秦颖. (2018). 基于神经网络的机器翻译质量评析及对翻译教学的影响. 外语电化教学(02), 51-56.
10. Sahari, Y., Al-Kadi, A. M. T., & Ali, J. K. M. (2023). A Cross Sectional Study of ChatGPT in Translation: Magnitude of Use, Attitudes, and Uncertainties. Journal of Psycholinguistic Research, 52(6), 2937-2954.
11. Siu, S. C. (2023). Chatgpt and GPT-4 for professional translators: Exploring the potential of large language models in translation. Available at SSRN 4448091.
12. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers (pp. 223-231).
13. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.

张雪晴

研究生，外国语学院，

哈尔滨工程大学，哈尔滨，中国

梁红

教授

外国语学院，

哈尔滨工程大学，哈尔滨，中国

跨模型比较：大语言模型在政治文本翻译中的效能分析

摘要: *随着大型语言模型（LLM）技术的飞速进步，其在包括自然语言处理和机器翻译在内的众多领域展现出了巨大的应用潜力。本文构建涵盖 200 个代表性政治文本的中译英测试集，将几种主流大型语言模型（包括 ChatGPT、Copilot 及文心一言）与 Google Translate 和 DeepL Translate 这两大主流翻译引擎生成的译文进行对比，旨在评估它们在政治文献翻译领域的性能。研究运用了 BLEU、CHRF++、TER 和 METEOR 四种自动评估工具来量化译文质量，并结合人工评价，以全面了解各个模型的性能。我们发现，ChatGPT 在与其他模型的对比中展示了一定的优势。然而，各模型在处理包含意识形态因素、复杂结构、特定文化词汇及隐喻等内容时仍具局限性，翻译准确性也有待加强。研究结尾探讨了在人工智能时代翻译任务的未来走向，并指出翻译工作者在机器翻译时代的角色不仅未被边缘化，反而变得更加重要。通过有效地整合人工智能技术，可以极大地提高翻译的质量和效率，同时促进跨文化交流和理解，为塑造中国的国际形象提供更全面的支持。*

关键词: *大语言模型；自动评估指标；政治文本翻译*