Lecture Notes on Language and Literature (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/langl.2024.070812
ISSN 2523-5869 Vol. 7 Num. 8

Examining the Reliability of Machine Translation in the AI Era: An Empirical Comparative Study of Four Translation Software

Lei Zhao[1,a,*], Jiajia Cao[1,b], Maojuan Lin[1,c]

[1]Faculty of Foreign Languages, Huaiyin Institute of Technology, Huai'an, China
[a]66264960@qq.com, [b]2665415158@qq.com, [c]3148547656@qq.com
[*]Corresponding author

Keywords: Machine translation, Artificial intelligence, Manual evaluation

Abstract: Currently, artificial intelligence (AI)-assisted foreign language translation has become a reality. However, human participation is still indispensable in distinguishing the advantages, disadvantages and reliability of various software. In this study, through the ranking method of manual evaluation, Chinese-to-English test translations were carried out on a total of four translation software of two types, namely Type A and Type B, from six aspects including words or phrases, proverbs, idioms, ambiguous sentences, political manuscripts and ancient Chinese poems. Based on the translation results, 10 teachers and 68 translation major students were selected to conduct online manual ranking. The ranking results show that the Type B artificial intelligence translation software is superior to the Type A online dictionary translation in the translation of words or phrases, proverbs and ambiguous sentences; however, it is not superior to the latter in the translation of idioms, political manuscripts and ancient Chinese poems, and none of them can accurately translate their Chinese connotations.

1. Introduction

In recent years, with the rapid development of machine translation, especially artificial intelligence technology, the evaluation of machine translation quality has become a hot topic in academic research [1]-[3]. Currently, the three mainstream methods for evaluating machine translation quality in the industry are: manual evaluation, automatic evaluation with references, and automatic estimation without references. "Each of the three methods has its own advantages and disadvantages, and the evaluation method should be flexibly selected according to different needs in specific scenarios" [2] 135. In addition, some scholars have systematically expounded the concepts, structures, characteristics, and advantages of the MQM (Multidimentional Quality Metrics) multi-dimensional quality index system [1]; however, its drawback lies in that "the evaluation process is time-consuming and laborious, and it requires relatively high capabilities of the

evaluators" [2] 137. Given the current lack of empirical studies based on the above evaluation methods, this paper takes the ranking method in manual evaluation as the evaluation criterion and compares the pros and cons of the translation quality of four translation software, hoping to provide references for improving the quality and reliability of machine translation.

2. Literature Review

2.1. Manual Evaluation: Ranking Method

Manual evaluation refers to the manual assessment of machine translation outputs based on certain evaluation criteria or indicators. The evaluation methods can be divided into two types: "scoring method" and "ranking method" [2] 136. As the name implies, the ranking method is to rank the translation results according to their quality levels. It is mainly used for horizontal comparison and provides a macroscopic estimate of different machine translations of the same source sentence. In the article "Research on the Standardization of the Evaluation of the Capability Level of Artificial Intelligence Machine Translation" [4], the author scored the quality of machine translation from six aspects, namely, translation fidelity, translation fluency, comprehensive error rate, pragmatic compliance, cultural compliance, and emotional compliance, and the results were relatively accurate and standard. Different from the scoring method, the ranking method does not need to give an exact evaluation of the translation quality but only provides a rough ranking of the quality levels. Therefore, it is not only highly efficient but also has a relatively high consistency of evaluation results. "The manual evaluation of the quality of machine translation outputs also has different evaluation characteristics due to different evaluators" [5] 77. Overall, the ranking method is simple to operate, time-saving and labor-saving, and has a relatively high ranking reliability, which is in line with this research.

2.2. Machine Translation

From a macroscopic perspective, current domestic research on machine translation mainly focuses on two categories, namely, the problems brought to the translation field by the era of artificial intelligence and large language models and their solutions. In the article "Problems and Reflections Caused by Machine Translation in the Digital Age" [6] 44, the author holds that:

Machine translation will bring a series of problems and impacts on foreign language education, the language life of the general public, and translators. For example, learners may lose their motivation to learn, their language and innovative thinking abilities may decline, application norms and ethical issues may be triggered, language diversity may be damaged, readers' enthusiasm for reading may be dampened, the awareness of the source language culture and way of thinking may be weakened, translation practitioners may be impacted, the subjectivity of translators may be weakened, translators' work enthusiasm and job satisfaction may decline, and people may easily become dependent on machine translation.

Secondly, there are empirical comparative studies on translation software, but they are relatively scarce. One study [7] 60 compared the translations of the novel "Flowers in Bloom" by multiple translation software. The results showed that "Wenxiaoyan and iFlytek Spark performed better than traditional domestic and foreign machine translation systems and ChatGPT in the Chinese-to-English translation of literary texts." Its shortcoming lies in that it only focused on one

type of literary translation and did not cover other literary types such as poetry, prose, and classical Chinese, lacking universal practicality. Undoubtedly, in the era of machine translation, the innovativeness and speculative nature of human translation should not be obliterated. How to make full use of the advantages of machine translation and avoid its disadvantages is a problem that the contemporary translation field needs to consider.

3. Research Design

This comparative analysis focuses on four translation software programs that are frequently used in China. They are divided into two categories: Category A and Category B. Category A consists of online electronic dictionaries, including Baidu Translator and Youdao Translator. Category B includes artificial intelligence software, such as Wenxiaoyan and Doubao Translators. The specific classification is shown in Table 1.

Table 1: Basic Information of the Four Software Programs in Categories A and B

Category	Software Name	Developing Company	Development Time
Category A	Baidu	Baidu	2011.6
	Youdao	NetEase	2008.8
Category B	Wenxiaoyan	Baidu	2023.3
	Doubao	TikTok	2023.6

The languages involved in this translation test include both Chinese and English, covering words, idioms, phrases, and sentences. The selection of Chinese materials generally has the following characteristics (as shown in Figure 1): The selection of words and phrases includes Internet buzzwords, local dialects, jargon, etc.; the selection of proverbs or idioms includes those commonly used in China, many of which have cultural stories behind them; the selection of sentences includes ambiguous sentences, ancient Chinese poems, etc. The translation of ancient Chinese poems is a major difficulty in the translation of literary works, as it is hard for machine translation to capture the emotions and contexts of the poems. In this test, "Thoughts in the Silent Night" by Li Bai is selected as the material of ancient Chinese poems to verify the reliability and accuracy of machine translation.

Category	Examples
Words and phrases	hāo yáng máo, qiáng tóu cǎo, bā shì, tuō gōu duàn liàn
Idioms	nìng wéi jī tóu, bù wéi fèng wěi; dāo zi zuǐ, dòu fu xīn; shī zhī dōng yú, shōu zhī sāng yú
Proverbs	bān mén nòng fǔ; sān gù máo lú; bī shàng liáng shān
Ambiguous sentences	guó pīng shuí yě yíng bù liǎo; guó zú shuí yě yíng bù liǎo; rè ài rén mín de zǒng lǐ
Political texts	zhuān jīng tè xīn xiǎo jù rén; móu huà quán jú xìng, fāng xiàng xìng, zhàn luè xìng wèn tí
Ancient Chinese poem	jìng yè sī lǐ bái chuáng míng yuè guāng, yí dì shàng shuāng. jǔ tóu wàng míng yuè, dī tóu sī gù xiāng.

Figure 1: Materials for Chinese-to-English Translation

Overall, the Chinese materials cover a wide range, including words, phrases, proverbs, idioms, ambiguous sentences, and ancient Chinese poems, which pose certain difficulties in translation. It requires a certain level of cultural background knowledge and proficiency in both languages. Then, we classify and archive the translation results and conduct manual evaluation to rank each software. Given the reliability of manual evaluation, the evaluators in this manual evaluation consist of 10 teachers from the front line of English teaching and 68 junior-year undergraduate students majoring in translation. Among them, all the teachers have more than five years of teaching experience and hold a master's degree or above. Based on the ranking results, we analyze and compare the advantages and disadvantages of the translation quality of the four software programs to provide references for artificial intelligence-assisted foreign language translation.

4. Results and Discussion

In terms of the ranking of Chinese-to-English translations based on the voting results of teachers and students, it is roughly as follows: From the perspective of inter-group comparison, the artificial intelligence translations in Category B are superior to the online dictionaries in Category A, and the voting results of teachers and students are roughly the same. Among them, the voting results of teachers in descending order are (see Table 2): Doubao (10 votes), Wenxiaoyan (9 votes), Youdao Translator (6 votes), and Baidu Translator (5 votes) respectively; the voting results of students are: Doubao (67 votes), Wenxiaoyan (63 votes), Youdao Translator (53 votes), and Baidu Translator (47 votes) respectively.

Table 2: Summary of Manual Evaluation Results

Category	Software	Teacher		Student	
		Num. Of Vote	Proportion	Num. Of Vote	Proportion
Category A	Baidu	5	50%	47	69%
	Youdao	6	60%	53	78%
Category B	Wenxiaoyan	9	90%	63	93%
	Doubao	10	100%	67	98%

Next, we will conduct a within-group and between-group comparative analysis of the Chinese-to-English translations from the aspects of words, sentences, and then ancient Chinese poems. Firstly, from the within-group perspective, in Category A, Youdao Translator is slightly better than Baidu Translate; in Category B, Doubao Translator is better than Wenxiaoyan Translator.

4.1. Word and Phrase Translation

In terms of words and phrases, Baidu Translate and Youdao Translate have misinterpretations and ambiguities in the translation of some individual words. “hāo yáng máo” is an Internet buzzword, which means "the behavior of users selectively participating in activities on various merchants and online platforms to obtain property benefits at a low cost or even zero cost". According to different purposes, “hāo yáng máo” can be divided into two categories: ordinary consumers and professional “wool parties”. Then according to these definitions, "Get the best deal" and "take advantage of loopholes or benefits" are the most appropriate translations, and users can

choose either one according to the actual context. Obviously, "plucking wool" is just its literal meaning and fails to achieve the purpose of translation, lacking accuracy. “qiáng tóu cǎo” is often used to metaphorically describe people who are not firm in their stance and have no personal judgment. Wenxiaoyan just gave an explanation, while both "fence-sitter" and "opportunist" can express this meaning. In addition, "weathercock" (weather vane) in English can also express the meaning of “qiáng tóu cǎo”.

The word “ba shi” belongs to the Sichuan local dialect, which means comfort and ease on the one hand, and suitability and just-rightness on the other hand. The translations of Category A online dictionaries are obviously behind those of Category B artificial intelligence translation software. Here, Wenxiaoyan's translation is more comprehensive and accurate than Doubao's translation. The term “tuō gōu duàn liàn”, decoupling and severing supply chains) was originally mainly used to describe the physical connection state. In the past two years, it has been frequently mentioned in the news reports about China-US and China-Europe relations. The English version of China Daily uses the phrase "decoupling and severing supply chains" translated by Doubao (see Example 1 and 2).

Example 1:...jointly opposing technology blockade and rejecting decoupling or severing supply chains,”Wang said. (China daily, 2024-09-29)

Example 2:...there are those who advocate “decoupling and severing supply chains”. (China daily, 2024-05-19)

Overall, the translations of Category B are more accurate than those of Category A. In Category A, Youdao Translation and Baidu Translation are basically on a par, each with its own shortcomings; in Category B, Doubao Translation is better than Wenxiaoyan Translation as it provides multiple translations for translators to choose from.

4.2. Proverb Translation

According to the social functions of language, the academic community roughly divides proverbs into two major categories: "proverbs that take imparting experience as the starting point and those that aim at enlightening life or implementing moral education" [8] 29, and proposes three methods for proverb translation: target language docking with the source language, source language adapting to the target language, and source language implantation [6]. Through comparison, the proverb "Better be the head of a chicken than the tail of a phoenix" reflects people's principle of doing things and is an expression of values. All four translations adopted the method of "source language implantation" to solve this. In addition, through searching in the Corpus of Contemporary American English (COCA), it was found that "Better be the head of a dog/cat than the tail of a lion" can also express this meaning. The proverb “dāo zi zuǐ, dòu fǔ xīn” means that someone speaks harshly and severely like a knife, but has a soft heart like tofu. The translations of all four software programs are acceptable. Only that Category A adopted the method of "target language docking with the source language" while Category B adopted the method of "source language adapting to the target language", and the expressive effects are the same. The proverb “shī zhī dōng yú,shōu zhī sāng yú” is often used to metaphorically mean that one fails in one place but succeeds in another, or suffers a loss in one aspect but gains in another. All four translations adopted the method of "target language docking with the source language", which is easy to understand.

Overall, the English translations of proverbs by the four software programs are all relatively ideal and have a higher accuracy rate than the translations of the above words and phrases. The

reasons for this are roughly as follows. First, some words are Internet buzzwords, such as “hāo yáng máo” (taking advantage of deals or benefits), “tuō gōu duàn liàn” (decoupling and severing supply chains), etc. They have emerged for a relatively short time, and their application contexts and scenarios are significantly different from previous expressions, resulting in Category A online dictionaries not having enough time to update, so there are phenomena such as misinterpretations and ambiguities. Second, proverbs are part of Chinese culture. They are the experiences, lessons, etc. summed up by the Chinese people during their long-term life, study, and tempering. They have existed for a long time and have a long history of use. Therefore, all four software programs can translate them relatively well.

4.3. Idiom Translation

Idioms are fixed phrases that have been formed through long-term use and refinement in Chinese. Idioms generally have clear origins, such as historical stories, fables, and legends. The translation of the idiom “bān mén nòng fǔ” (Showing off one's proficiency with the axe before Lu Ban) is only relatively accurate by Doubao Translator. It not only translated the figure (Lu Ban) behind it but also gave the practical meaning it expresses. While the other three translations only translated its practical meaning. The idiom “sān gù máo lú” (Visiting the thatched cottage three times) comes from The Romance of the Three Kingdoms and expresses the spirit of being eager to seek the worthy talent. To understand the translations of Baidu and Doubao, one needs to know its origin or the background knowledge of Chinese culture; otherwise, it is easy to misunderstand. While Youdao and Wenxiaoyan both mentioned the purpose of seeking the worthy and entrusting them with important tasks, which is relatively accurate. The idiom “bī shàng liáng shān” (Being driven to revolt on Liangshan) comes from Water Margin and expresses the social situation of political corruption, the dominance of treacherous officials, and the cruel exploitation of the people by the rulers during the Northern Song Dynasty. It describes that under various oppressions and difficulties, people are forced to take extreme or rebellious actions. The translations of all four are acceptable, using words such as "be driven", "be forced", "revolt", "desperate", "take refuge", etc. Overall, the translations of idioms by the four software programs are relatively accurate, with high reliability and accuracy rates.

4.4. Ambiguous Sentence Translation

Ambiguous sentences refer to those that may have two or more possible meanings in understanding. Take the example of "The Chinese national table tennis team can't be beaten by anyone; The Chinese national football team can't beat anyone". The Chinese national table tennis team has won honors for the country in various major world competitions, while the Chinese national football team has been in a slump and has lost repeatedly. Although both sentences use "can't... anyone", their meanings are completely opposite. The online dictionaries in Category A did not accurately translate the meaning of this sentence; especially Baidu Translate, which completely misinterpreted it. While those in Category B could accurately express the meaning of the sentence, which also reflects the advantages of generative deep learning artificial intelligence. Here, Wenxiaoyan provided two different translation methods, such as the use of words like "defeat", "invincible", etc.

The sentence “rè ài rén mín de zǒng lǐ” has two meanings, namely, the premier who loves the

people and loving the premier of the people. But all four translations adopted the understanding of "the premier who loves the people". Even in the sentence “wǒ men rè ài rén mín de hǎo zǒng lǐ”, Doubao translated it as "We love our premier who loves the people". However, when translated back into Chinese, it becomes “wǒ men rè ài wǒ men rè ài de zǒng lǐ”.

In conclusion, we can see that artificial intelligence software still has shortcomings when translating ambiguous sentences and still requires translators to carefully distinguish. Of course, overall, Category B is still better than Category A.

4.5. Political Texts

Chinese emphasizes parataxis, while English emphasizes hypotaxis [9]. In the selected Chinese materials this time, each of “zhuān jīng tè xīn” represents different meanings, but all are used to modify enterprises like "little giants". Different from the machine translations of Category A and Category B, the English translation by the Ministry of Foreign Affairs of China is as follows: "‘little giant' enterprises that use special and sophisticated technologies to produce novel and unique products". It is not difficult to see that the translator handled the four adjectives very precisely. "Specialized" and "refined" modify "technology", while "unique" and "innovative" modify "products". Another material is “móu huà hǎo quán jú xìng、fāng xiàng xìng、zhàn luè xìng wèn tí”. None of the four software programs could accurately translate it, while human translation was better: "consider and devise our strategy to address issues that have a global impact, determine our future direction, and possess strategic significance". From the English translation, it can be seen that the translator accurately understood that the object of "planning" is "strategy", not "issues". At the same time, the purpose of "planning strategy" is to solve the "overall, directional, and strategic issues". In this way, the semantics are smooth, avoiding ambiguity and misunderstanding. Unfortunately, whether it is Category A or Category B, they can only do a rigid translation according to the literal meaning and did not accurately translate the real meanings of the two materials.

4.6. Ancient Chinese Poem Translation

Chinese classical poems are an indispensable and important part of Chinese literature. The “literariness” of classical poems is a common poetics concept in both China and the West [10] 95. How to embody the literariness of classical poems in translations has always been one of the hot topics in the translation field. "Based on the two dimensions of the poetic sense on the language level of the translation and the poetic meaning when the translation is an independent text, discussing the realization path of the literariness of the translations of Chinese classical poems" [10] 95 has certain practical guiding significance. However, after reading through the four translations in the following table, there is no “literariness” at all. Chinese classical poems have the characteristics of profound artistic conception, concise language, harmonious rhythm, and sincere emotions. Professor Xu Yuanchong continuously summarized the methods of translating rhymed poems on the basis of translation practice [11] 55, making important contributions to the "going out" of Chinese classical poems. Take Li Bai’s "Thoughts in the Silent Night" as an example. The poem describes what the poet saw and felt late at night when he was staying away from home, reflecting a strong sense of homesickness. The poem has a beautiful rhythm (see Figure 2 below), such as "Guang" (light), "Shuang" (frost), and "Xiang" (hometown) all rhyming at the end. By comparing the translation by Professor Xu Yuanchong, it can be seen that "Light" and "Bright", "Around" and

"Drowned" all rhyme, which is catchy and reproduces the original rhythm of the poem.

A Tranquil Night
Translated by Xu Yuanchong
A bed, I see a silver light,
I wonder if It's frost around.
Looking up, I find the moon bright,
Bowing, in homesickness I'm drowned.

Figure 2: A Tranquil Night translated by Xu Yuanchong

In conclusion, it is not difficult to see that neither online dictionaries nor artificial intelligence software can accurately reflect the beauty of Chinese classical poems. They can only translate the literal meanings, losing the rhythm and "literariness" of the poems, and thus are not worthy of recommendation for use.

5. Conclusion

This research conducted Chinese-to-English translation tests on a total of four translation software programs in two categories, A and B, from five aspects including words or phrases, proverbs, idioms, ambiguous sentences, and ancient Chinese poems through the ranking method of manual evaluation. According to the translation results, 10 teachers and 68 translation major students were selected to conduct online manual ranking. The ranking results showed that the artificial intelligence translation software in Category B was superior to the online dictionary translation in Category A in terms of word or phrase, proverb, and ambiguous sentence translations; however, it was not superior to the latter in idiom, political text, and ancient Chinese poem translations and could not accurately translate their Chinese meanings. Undoubtedly, talent cultivation is the main theme of education. In the era of artificial intelligence, we should uphold integrity and innovation, enable artificial intelligence to empower foreign language teaching, and place more emphasis on the cultivation of "human beings" [12]-[14]. This research also has its shortcomings. For example, only some materials for Chinese-to-English translation were selected, and the actual effects of the four software programs in English-to-Chinese translation were not tested; secondly, the ranking method of manual evaluation is not as accurate as the scoring method, only roughly counting the degrees of superiority and inferiority of each software program, which remains to be discussed.

Acknowledgements

1) General Project of Philosophy in Colleges and Universities in Jiangsu Province: "A Study on Deliberate Metaphors from the Perspective of New Cognitive Pragmatics" (2021SJA1824);

2) School-level Teaching Reform Project of Huaiyin Institute of Technology: Research on the Teaching Practice of Reflective Writing in the Classroom of "Teacher-Student Collaborative Assessment".

References

[1] Tian Peng. An Introduction and Evaluation of the MQM Model of Multidimensional Translation Quality Standards

and Its Implications [J]. Eastern Journal of Translation, 2020(03):23-30.
[2] Wang Junsong, Zhuang Congqian, Wei Yongpeng. Machine Translation Quality Evaluation: Methods, Applications and Prospects [J]. Foreign Languages and Literatures (Bimonthly), 2024, 40(3):135-144.
[3] Yu Lei. A Study on the Lexical Diversity and Syntactic Complexity of ChatGPT Translations [J]. Foreign Language Teaching and Research (Bimonthly), 2024, 56(2):297-307.
[4] Ma Wanzhong, Liu Junhua, Zhu Xiaojie. Research on the Standardization of the Evaluation of the Capability Level of Artificial Intelligence Machine Translation [J]. Information Technology and Standardization, 2020(Z1):21-26.
[5] Ma Minghao. A Preliminary Discussion on Machine Translation Quality Evaluation [J]. Journal of Ningbo Institute of Education, 2019, 21(06):76-78+93.
[6] Zhao Yan, Zhang Hui, Yang Yi. A Comparative Study on the Quality of Text Translations by Large Language Models—Taking the Translation of "Flowers in Bloom" as an Example [J]. Computer-Assisted Foreign Language Education, 2024(04):70-78.
[7] Shan Xinrong, Xiao Kunxue. A Study on the Translation Strategies of Figurative Proverbs from the Perspective of Metaphor Comprehension [J]. Foreign Language Education Research, 2024, 12(03):29-34.
[8] Lian Shuneng. 1993. A Comparative Study of English and Chinese[M]. Beijing: Higher Education Press.
[9] Zhou Fangheng. A Study on the Realization Path of the Literariness of Translations of Chinese Classical Poems [J]. Journal of Xi'an International Studies University, 2023, 31(4):95-100.
[10] Li Zhengshuan, L ü Xin. A Study on the Translator's Behavior in the Chinese Translation of Burns' Poems by Wu Fangji [J]. Journal of Xi'an International Studies University, 2024, 32(02):55-60+95.
[11] Hu Jiasheng, Qi Yajuan. Foreign Language Education in China in the ChatGPT Era: Seeking Changes and Adapting to Changes [J]. Computer-Assisted Foreign [11] Language Education, 2023(01):3-6+105.
[12] Liu Lei, Zhang Xinya. Reflections on the Impact of Artificial Intelligence Dependence on Creativity and the Future Development Path of Education [J]. Journal of Guangxi Normal University (Philosophy and Social Sciences Edition), 2024, 60(1):83-91.
[13] Li Hanji, Chen Haiqing. Research on Translation Ethics in the Technological Era: Challenges and Expansions [J]. Journal of Northeastern University (Social Sciences Edition), 2024, 22(1):112-119.
[14] Wang Zhenzhen. Problems and Reflections Caused by Machine Translation in the Digital Age [J]. Journal of Yunnan Normal University (Philosophy and Social Sciences Edition), 2024, 56(5):44-54.