Original Paper

Qualities of Literary Machine Translation: A Corpus-based Case Study

Tian Jiaxin[1*]

[1] Chongqing University of Posts and Telecommunications, Chongqing, China

[*] Tian Jiaxin, Chongqing University of Posts and Telecommunications, Chongqing, China

Received: June 15, 2024 Accepted: June 30, 2024 Online Published: July 24, 2024

doi:10.22158/sll.v8n3p39 URL: http://dx.doi.org/10.22158/sll.v8n3p39

Abstract

The research assesses and compares the translation performance of two popular machine translation systems, GPT-4o and Youdao AI Translate, in translating into English ten Chinese prose essays excerpted from Selected Modern Chinese Essays 2 by Zhang Peiji. The goal is to discover their linguistic features and investigate how well they can perform in this translation. Through a corpus-based analysis, the research explores the STTR and word/ sentence length of their translations and conducts both automated and human evaluations on their translation quality. It reveals that GPT-4o exhibits higher lexical variety and both the two machine translation systems tend to produce more and shorter sentences than the human translation does. Both of them perform surprisingly well in the translation, as they get relatively high BLUE scores yet low TER scores, as well as high adequacy and fluency rates. Our evaluation results also show that Youdao AI Translate displays generally better performance than GPT-4o in the translation of Chinese-to-English literary texts, and they can complement each other to achieve even better performance, though a certain amount of errors are still present in both of their translations.

Keywords

machine translation, ChatGPT, corpus, translation quality, literary translation

1. Introduction

Machine Translation (hereafter MT) has experienced three types of paradigms in its development. The first is rule-based machine translation (RBMT) which uses bilingual dictionaries and manually written rules to translate source language texts into target language texts (Haifeng Wang et al., 2022). Since RBMT is labor-intensive, corpus-based systems, which employ bilingual corpora of parallel texts to create translations (Hutchins, 1995), began to gain in popularity at the end of the 1980s (Hutchins, 2007). These systems can be divided into example-based systems (EBMT) and statistical MT systems (SMT). By the 2010s the quality of SMT output in well-supported language pairs had plateaued (Rothwell, A. et al., 2023), and SMT served as the predominant paradigm until the emergence of neural machine translation (NMT) in 2014. Compared with previous methods such as RBMT and SMT, NMT does not need human-designed rules and features. NMT is an end-to-end framework that directly learns semantic representation and translation knowledge from the training corpora (Haifeng Wang et al., 2022). NMT is currently the most popular and widely used method in the MT community, particularly for the translation of non-literary texts. MT can now produce surprisingly good output in many language pairs, especially between languages that are closely related, making it useful for both translators and end users (Rothwell, A. et al., 2023).

As for literary texts which are considered perhaps the most challenging tasks for NMT, it still remains debatable, however, as to whether NMT is capable of generating readable and quality literary translations. Rothwell (2023, p. 1) believed that while certain aspects of technology can be argued to have the potential to negatively affect the work of literary translators, this is by no means true of all translation technologies, which collectively offer a range of creative opportunities to enhance their work.

In this article, we assess and compare the Chinese-English translation performance of ChatGPT-4o and Youdao AI Translate through a corpus-based methodology. ChatGPT is an intelligent chatting machine developed by OpenAI upon the InstructGPT (Ouyang et al., 2022), and it is currently attracting great interest (Rothwell, A. et al., 2023). GPT-4o is the newest flagship model, at the time of writing, that provides GPT-4-level intelligence but is much faster and improves on its capabilities across text, voice, and vision. Youdao AI Translate is a popular MT system developed by Netease, a company headquartered in China, with a large share in the Chinese translation market.

We select ten articles excerpted from *Selected Modern Chinese Essays 2* by Zhang Peiji as the source texts and compare the English translations by GPT4o and Youdao AI Translate with the human translations by Zhang Peiji, a renowned Chinese translator, through the lens of corpus-based analysis and automatic and human evaluation results. Therefore, this article delves into the following research questions: (1) What are the linguistic features of NMT, compared with those of HT, in the translation of literary texts, or Chinese prose essays in particular in this study? (2) Can NMT be effectively utilized and trusted in Chinese-to-English literary translation?

2. Literature Review

A growing number of studies have been conducted on the machine translation of literary texts in recent years, most of which focused on the machine translation between English and other languages.

Thai, K. et al. (2022) collected a dataset of human and automatic English translations of non-English language novels. They discovered that expert literary translators preferred reference human translations over machine-translated paragraphs at a rate of 84%, and MT outputs contained not only mistranslations, but also discourse-disrupting errors and stylistic inconsistencies.

Macken, L. et al. (2022) examined three different versions of English-Dutch literary translations: the DeepL output, the post-edited version and the revised translation. MT quality was in line with expectations with 27.5% error-free sentences. The three main error types were various kinds of mistranslations, disfluent sentence constructions and different types of spelling and punctuation problems. The results show that more editing occurred during revision than during post-editing with different types of editing actions.

Fakih, A. et al. (2024) investigated the potential of Neural Machine Translation utilized by Instagram in producing high-quality translations from Arabic to English. From 30 selected Instagram captions with literary content, the study found Instagram’s machine translation lacking in 90% of cases, particularly in accuracy, fluency, and style with 61 errors were identified, adversely affecting the quality and failing to convey the original message.

As for machine translation between English and Chinese, most of the studies focused on translating English literary texts into Chinese.

Hu and Li (2023) compared the Chinese translations of Shakespeare’s plays conducted via DeepL to the translations by Liang Shiqiu. The study found that DeepL performed well in translating these works, with an accuracy and fluency rate of above 80% in sampled texts, showing the potential of the use of neural machine translation in translating literary texts across distant languages. The research further revealed that the DeepL translations exhibited a certain degree of creativity in their use of translation methods, even though there were still a certain amount of translation errors including literal translations.

He et al. (2024) compared two machine translation systems, Bing Translator and Youdao Machine Translation, using selected texts from the novel “Nineteen eighty-four” by George Orwell. The study revealed that Youdao Machine Translation demonstrated superior performance in accuracy while Bing Translator yielded more fluent and natural-sounding translations. The study highlighted the indispensable of human post-editing in achieving optimal accuracy, fluency, and overall readability in literary translations.

Studies retrieved above show that MT has been applied to literary translation, though there exist many errors of different types in the translated texts. Most of the studies concern the machine translation between English and some closely-related languages, instead of some distant languages such as English and Chinese. Few studies pay attentions to the Chinese-to-English machine translation of literary texts. And it remains unclear about the linguistic features of machine translated English literary texts, and whether NMT systems trained with more English corpus would outperform those with more Chinese corpus in the Chinese-to-English translation of literary texts.

3. Research Design

3.1 Corpora Compilation

With the aim to explore the performance of NMT in the Chinese to English literary translation, we select ten essays, randomly from *Selected Modern Chinese Essays 2* as the source texts. Three corpora are established, which are composed of the original ten essays in Chinese, and their corresponding translations by GPT-4o and Youdao AI Translate, as well as the translations by Zhang Peiji. A prose essay, as opposed to verse or poetry, is a type of essay in a natural, flowing form of language. As said by Zhang Peiji (2007), prose essay is flexible in style with straightforward discourse. China has a long history of prose essay writing and has yielded great achievements comparable to poetry. But most of the Chinese literary works translated and introduced to the outer world are mainly poems and novels, instead of prose essays. That is the reason for us to choose prose essays and its English translations as the corpora for research.

To get the translations by GPT-4o and Youdao AI Translate, we firstly use the unified prompt of “Please translate the following essay into English + the source text” to get responses from GPT-40. It is noted that only the first response of each essay from GPT-4o was recorded and incorporated into the corpus, with no follow-up prompts sent to further polish its translation. As for Youdao AI Translate, we just provide the source texts to it and no prompt is required. Also only the first translation version of each essay is cited into the corpus, even though we could use follow-up prompts for the system to further revise its translations for personal preference.

3.2 MT Quality Evaluation

As there is no single definitive standard to measure translation quality, in this work we combine both automated and human metrics to assess the quality of translations by GPT-4o and Youdao AI Translate.

The MT community extensively uses automatic metrics that score candidate translations against references. Two widely-used automated metrics are thus utilized in this research to evaluate MT quality. Bilingual Evaluation Understudy (BLEU, Papineni et al. 2002) is the de facto standard metric in the MT field (Antonio Toral et al., 2015). BLEU measures both adequacy by looking at word precision and fluency by calculating n-gram precision for n =1,2,3,4 (Lavie, A. 2021). It has become very popular by showing good correlation with human judgement. While Translation Error Rate (TER) measures the amount of editing that a human would have to perform to change a system output so it exactly matches a reference translation. (Snover et al., 2006). The editing operations include shift, substitution, deletion, and insertion. The TER metric is expressed as the ratio of the number of edits to the average number of words in the reference (Lee, S. et al., 2023). Though automated metrics save time and labor in evaluation, there exist shortcomings and limits. BLEU and TER are both lexical-based metrics. BLEU only considers exact word matching and it does not reflect the meaning and sentence structure of the translation result (Lee, S. et al., 2023). TER ignores the semantic similarity, and a semantically correct translation may still gets a bad score if it’s words do not match with the reference translation. Therefore, automated metrics can only reflect a certain sides of translation quality, and human evaluation is necessary, in order to realize comprehensive MT quality evaluation.

As for human evaluation, we manually assess the performance of MT by identifying adequacy and fluency errors. Adequacy evaluates semantic quality, that is, if the information has been correctly transmitted or not, which requires comparison with with the original text (Irene Rivera-Trigueros, 2021). While fluency evaluates syntactic quality (Irene Rivera-Trigueros, 2021), which instead focuses on the target text and is typically defined as the extent to which the translation follows the rules and norms of the target-language (regardless of the source or input text) (Sheila Castilho et al., 2018). The combined utilization of both automated and human metrics is conducive to reaching a more well-round evaluation of machine translation quality.

3.3 Research Procedures

To explore the features of translations by GPT-4o and Youdao AI Translate and assess their performance in Chinese-English literary translation, we conduct three-step analysis. Firstly, we use WordSmith 9.0 software to analyze some corpus-based features of NMT translations, including the type/token ratio and mean word/ sentence length. The aim is to answer the first research question we raised. What are the linguistic features of NMT? Secondly, we turn to automated MT quality evaluation metrics and compute the BLEU and TER scores of translations by the two NMT systems. Detailed comparison and analysis are carried out to demonstrate their performance. And finally, human evaluation is also conducted from the aspects of adequacy and fluency to better assess the NMT translations from another perspective, so as to provide insights to the second research question: Can NMT be effectively utilized and trusted in Chinese-English literary translation?

4. Results and Discussion

4.1 Linguistic Features of NMT

In this section, we conduct through WordSmith 9.0 a corpus-based analysis of the translations by GPT-4o (hereinafter GT) and by Youdao AI Translate (hereinafter YT) as well as by Zhang Peiji (hereinafter ZT) in aspects of their type/token ratio, mean word length and mean sentence length, to figure out the features of these two popular NMT systems against human translations. Concrete translation cases of the three versions are provided and compared to demonstrate their different translation preferences.

4.1.1 Type/Token Ratio

Table 1 reveals that ZT and YT are very close in the amount of tokens (8,886 and 8,975), while GT contains far less tokens (8,046), making its STTR the highest among the three (44.10), followed by ZT(43.13) and YT (40.02). We can see that GT has the outstanding performance in lexical variety, even surpassing the human translations. YT, however, is relatively lower in lexical richness, which may partly attribute to its smaller English training corpus compared with that of GPT-4o.

Table 1. Type/Token Ratio

Files	Tokens	Types	TTR	STTR	STTR Std. dev.	STTR basis
ZT	8,886	2,334	26.27	43.13	45.68	1,000
GT	8,046	2,283	28.37	44.10	44.46	1,000
YT	8,975	2,173	24.21	40.02	49.77	1,000

4.1.2 Mean Word/Sentence Length

It can be learned from Table 2 that GT generally prefers longer words than ZT and YT do. As to sentence length, Both GT and YT contains far more sentences (427 and 574) than ZT (365) does, and their average sentence lengths (17.85 and 15.05 in words) are thus relatively shorter than that of ZT (23.15). It indicates that in this case both the two NMT systems produce more and generally shorter sentences. As there are many short Chinese sentences in the source text, which is a distinctive feature of Chinese, particularly in Chinese literary works, NMT systems may follow closer the source text to produce short sentences, while human translations always render several Chinese short sentences into a long English one.

(1) **ST**:韩愈援引孔子的先例，作出判断说，”圣人无常师”。这句话的意思是说真正聪明有学问的人，没有一定的老师；见人有学问，不管是谁，就认他为师。

ZT: Han Yu, going by Confucius' teaching, asserts that "a sage has no definite teacher", meaning that a really wise and learned person has no fixed teacher and that he learns from whoever knows.

GT: Han Yu cites Confucius's example and makes a judgment, saying, "The sage has no constant teacher." The meaning of this sentence is that truly intelligent and knowledgeable people do not necessarily have a fixed teacher; when encountering someone knowledgeable, regardless of who they are, one should regard them as a teacher.

YT: Han Yu cited the precedent of Confucius and concluded, "A sage has no constant teacher". The meaning of this statement is that truly intelligent and learned people have no fixed teachers; whenever they meet people with knowledge, regardless of who they are, they recognize them as teachers.

In example (1), ZT just reorganizes the three original Chinese sentences into a single English one to make it more concise and fluent, while GT and YT largely stick to ST and produce three word-for-word long sentences. In this sense, NMT systems are obviously still falling behind HT, though their translations are truly readable and understandable.

Nevertheless, it doesn't mean that NMT systems completely follow ST without any revisions. They indeed, on the contrary, have good performance in correctly dividing the original long Chinese sentence into several ones with different sentence patterns, which is also a translation technique utilized by human translators, as shown by example (2).

(2) **ST**: 大家团团围着矮凳坐下，点了两壶龙井茶，和一些奇甜的东洋点心，她们吃着喝着，高声谈笑着，她们真像是才出谷的雏莺，只觉眼前的东西，件件新鲜，处处都富有生趣。

ZT: They sat around a small table and ordered two potfuls of longjing tea and some extremely sweet Japanese pastries. They chatted and giggled loudly over the snack, like young orioles just fledged. They found novelty in everything before them and joy of life here and there.

GT: Sitting around low stools, they ordered two pots of Longjing tea and some peculiarly sweet Eastern sweets. Eating and drinking, they chatted and laughed loudly, feeling as carefree as young birds just leaving the nest, finding everything before their eyes fresh and full of life.

YT: They sat around the low stools in a group, ordered two pots of Longjing tea and some extremely sweet Japanese snacks. They ate, drank and talked loudly. They were truly like newly hatched orioles, finding everything before them novel and full of vitality.

Table 2. Mean Word/Sentence Length

Files	Mean word length	Word length std.dev.	Sentences	Mean (in words)	Std.dev.
ZT	4.50	2.42	365	23.15	25.21
GT	4.74	2.48	427	17.85	9.71
YT	4.54	2.40	574	15.05	8.12

4.2 Automated Evaluation

The BLEU and TER scores of GT and YT are computed through python, with ZT as the reference translation. The results are displayed in Table 3.

BLEU scores can be divided according to Table 4 into seven ranges, each indicating a different level of translation quality compared to the reference translation. The BLEU ranges from 0 to 1, and the higher the score, the better the translation. We can see that both GT and YT are rated high BLEU scores, mostly higher than 0.5, in the ten essay translations. The average BLEU score of GT hits 0.54, which refers to very high quality, adequate, and fluent translations, while YT gets a even higher average score (0.60), reaching the highest level of quality often better than humans. To look into the BLEU score of each essay, we can find that YT gets a better score than GT in each essay and YT has a score over 0.5 (6 between 0.50-0.59, 3 between 0.60-0.69 and 1 over 0.7) in each translation, while GT has 3 translations between 0.40-0.49, 4 between 0.50-0.59 and 3 over 0.60. As a result, Both GT and YT generally offer high-quality translation and get an overall excellent BLEU score in each essay translation, with YT slightly higher than GT in both the average score and scores in each task. While BLEU is a widely used metric, it has its limitations. It focuses on n-gram precision and does not account for semantic equivalence, fluency, or the overall coherence of translations. Therefore, a high BLEU score does not necessarily mean the translation is flawless in all aspects, and TER score is often provided as a complement.

There is not a definite standard interpretation of TER score, though, a higher TER score indicates more-post editing effort and the lower the score, the better the translation, as it requires less post-editing. In this case, both GT and YT get relatively high TER scores, all higher than 0.50, but the average TER scores reflect the same result as the average BLEU scores do, that is, YT, with an average of 0.66 in TER score, shows comparatively better translation performance than GT, whose average TER score is 0.70. To check the score of each essay, we can see that eight out ten of translations by YT get lower TER scores than those by GT, with the scores of the other two (0.72 and 0.79) just slightly higher than those by GT (0.66 and 0.78). From the statistics, it can be inferred that both GT and YT are rated high TER scores, requiring much post-editing effort to align their translations with reference translation. But taking both the two metrics into consideration, we may safely arrive at the conclusion that YT achieves better performance than GT.

The BLEU score and TER score also show clear correlation with each other. High BLUE scores are always linked with low TER scores and vice versa. For instance, the translations of essay 3 by GT and YT both receive the highest BLEU score (0.65 and 0.72) yet the second lowest or the lowest TER score (0.54 and 0.52) among their ten translations, while the translations of essay 9 by GT and YT both receive the second lowest BLEU score (0.46 and 0.53) yet the highest TER score (0.88 and 0.79) among their ten translations. Some segments from these two essays are extracted for further examination.

Table 3. Automated Evaluation Results

Essay No.	BLEU		TER	
	GT	YT	GT	YT
1	0.57	0.59	0.66	0.72
2	0.62	0.67	0.54	0.52
3	0.65	0.72	0.59	0.52
4	0.60	0.66	0.71	0.63
5	0.56	0.62	0.74	0.70
6	0.49	0.55	0.67	0.60
7	0.50	0.51	0.78	0.79
8	0.53	0.58	0.68	0.65
9	0.46	0.53	0.88	0.79
10	0.45	0.54	0.74	0.69
Average	0.54	0.60	0.70	0.66

Table 4. Interpretation of BLEU Scores

BLEU	Interpretation
＜ 0.1	Almost useless
0.1-0.19	Hard to get the gist
0.2-0.29	The gist is clear, but has significant grammatical errors
0.3-0.39	Understandable to good translations
0.4-0.49	High quality translations
0.5-0.59	Very high quality, adequate, and fluent translations
≥0.6	Quality often better than humans

(3) **ST**:日本军国主义的残余到今天还拒不承认日本帝国主义侵略中国和亚洲的事实，但作为一个中国人的我，却不能忘却日本军人对中国人民犯下的血债。--冯亦代《忘了过去就是犯罪》(From essay 3)

ZT: Today remnants of the Japanese militarists still refuse to admit the aggression against China and other Asian countries by Japanese imperialism. As a Chinese, however, I can never forget the blood debts owed by the Japanese military to the Chinese people.

GT: The remnants of Japanese militarism still refuse to acknowledge the fact that Japanese imperialism invaded China and Asia. But as a Chinese person, I cannot forget the blood debt owed by Japanese soldiers to the Chinese people. (BLUE 0.66, TER score 0.42)

YT: To this day, the remnants of Japanese militarism still stubbornly refuse to admit the fact that Japanese imperialism invaded China and Asia. But as a Chinese, I can never forget the blood debt that the Japanese military inflicted upon the Chinese people. (BLEU score 0.75, TER score 0.49)

(4) **ST**:不过这还是极少数”痴子”的信心，一般人还是嗤之以鼻。--邹韬奋《什么事不可能》(From essay 9)

ZT: Yet the handful of “dreamers” of those days were subjected to jeers and laughter.

GT: However, this is still the confidence of a very small number of “madmen”, and the general public still sneers at it. (BLUE 0.19, TER score 1.00)

YT: But this was only the belief of a very small number of “crazies”, and the majority of people still scoffed at it. (BLEU score 0.17, TER score 1.00)

Example 3 is a segment translated by GT and YT with high BLUE scores and low TER score. The two NMT systems both produce nearly perfect translations. One of the reasons may lies in that the ST is a fact-based historical essay and the language is plain and easy to understand. But when it comes to essay with personal emotion, the situation would be different and the automated metrics may not be that dependable. As evidenced by example 4 where both GT and YT get a very low BLEU scores yet a 1.00 TER score. That means the translation, though semantically correct, is completely different from the reference, which is actually the fact. This is evidenced by Omazić, M. et al. (2023), who argue that while automatic assessment methods offer significant advantages in terms of speed and cost, they cannot fully capture the nuances and context-specific aspects of human language. Therefore, human evaluation is often conducted in combination with these automated metrics to better shed light on MT quality.

4.3 Human Evaluation

As mentioned above, we detect manually errors affecting adequacy and fluency in the human evaluation section. We firstly finish the sentence-level alignment of the source texts with GT and YT. There are 357 Chinese sentences in total, which are rendered, as discussed above in the 4.1.2 section, into 427 and 574 English sentences by GT and YT respectively. The human evaluation is carried out based on each Chinese segment and its corresponding English translations. An adequacy /fluency error sign is marked on a segment as long as one kind of these errors is found in it, therefore, the number in Table 5 refers to how many segments of the translation contain errors, instead of how many errors are exactly detected in the translation since there may be more than one error in a segment.

As we can see in Table 5, both GT and YT achieves very surprisingly high adequacy /fluency rates in this task, but YT gets 90.80% in adequacy rate and 93.28% in fluency rate, which means that it still proves to have better performance than GT, whose rates are 86.83% and 85.15%. The results are in line with those of automated evaluation.

Table 5. Human Evaluation Results

	Adequacy			Fluency		
	Errors	Error rate	Adequacy rate	Errors	Error rate	Fluency rate
GT	47	13.17%	86.83%	53	14.85%	85.15%
YT	33	9.20%	90.80%	24	6.72%	93.28%
Number of sentences (in Chinese): 357						

For detailed analysis, GT commits more fluency errors, with two segments showing both adequacy and fluency errors. While YT contains more adequacy errors, with only one segments showing both kinds of errors. There are 18 segments translated with adequacy errors by both GT and YT, and 13 segments with fluency errors by both of them. That means some of the adequacy or fluency errors made by GT are actually avoided by YT and vice versa.

Adequacy errors are often associated with literal translation and omission, particularly the former, which is also evidenced by Hu and Li (2023) in their research of DeepL's English to Chinese translation quality. And that also applies to Chinese to English translation. NMT systems still tend to conduct word-for-word translation without taking the context into consideration. In example 5, “赤裸裸地” is just rendered into the ambiguous “nakedly” by both GT and YT, which actually means “fully” here in this context.

Fluency errors are often made in the translation of long sentences, mainly including unidiomatic expression, unnatural sentence structure and grammar mistakes. NMT systems occasionally fail to render the original long sentence into a natural and logically organized sentence in target language. They just follow the source text to produce some incomplete phrases. In example 6, GT and YT both follow the order of the original sentence to give a long yet not so fluent sentence.

(5) **ST**:朋友，在现在这一刻这一个地方，把你的青春的力，你的生命**赤裸裸地**表现出来呵！你要把你的生命变做狂风，变做暴雨，把世界上一切肮脏的东西扫荡一个干净。--张闻天《生命的急流》

ZT: Friends, bring your youthful vitality and life **into full play** right now and here! Turn your life into a hurricane to make a clean sweep of all filth on earth.

GT: Friend, at this moment, in this place, reveal your youthful strength, your life **nakedly**! Turn your life into a gale, into a storm, sweep away all the dirt in the world.

YT: My friend, at this moment and in this place, show your youthful strength and your life **nakedly**! You should turn your life into violent wind and heavy rain and sweep all the dirty things in the world clean.

(6) **ST**:在那枝柯掩映中，一道金黄色的柔光正荡漾着。使我想象到一个披着金绿柔发的仙女，正赤着足，踏着白云，从这里经过的情景。--庐隐《异国秋思》

ZT: A ray of golden sunshine gently filtering through the tree branches conjured up in me the vision of a golden-haired fairy maiden treading barefoot on white clouds on her way through the place.

GT: Amidst the branches and leaves, a golden soft light shimmered, conjuring images of a fairy with golden-green hair, barefoot, stepping on white clouds passing through.

YT: Among the intertwined branches and leaves, a golden-yellow soft light was undulating, making me envision a fairy with golden-green soft hair, barefoot and stepping on white clouds, passing by.

5. Conclusion

The research looks into the performance of NMT systems, GPT-4o and Youdao AI Translate in particular, in the Chinese to English translation of literary works, or actually prose essays. We have accomplished the goal by carrying out three stages of assessment: corpus-based analysis, and automated and human MT quality evaluation. Our study reveals that both GT and YT produces far more sentences than ZT does. GT has a relatively small amount of tokens and higher STTR, and YT contains more tokens with a quite lower STTR. That means GT has a higher lexical variety, though both GT and YT tend to generate shorter sentences compared with ZT. The quality evaluation believes that YT outperforms GT in the translation task measured by the automated metrics of BLEU and TER scores, as YT gets a higher average BLEU score (0.60) over GT (0.54) yet a lower average TER score (0.66) than GT (0.70). It indicates that YT is closer to ZT in similarity and less post-editing effort is required. What is found in the human evaluation also support the hypothesis, as YT contains relatively fewer errors concerning both adequacy and fluency. However, both GT and YT inevitably make mistakes in these two aspects, especially literal translation and unnatural sentence structure.

Based upon the analysis, we may come to the conclusion that NMT, with its rapid development, can be effectively utilized and trusted to a certain extent in Chinese-English translation of prose essays. It reveals that a NMT system trained with more corpora in source language tends to produce translations that better transfer the actual meaning of source texts into semantically correct translations, as it has a better understanding of source texts. But there are still some losses in its translation such as lower lexical variety and longer translation in length. Therefore, each NMT system has its own strength and an integrated utilization of more than one NMT system might be deemed as helpful in improving MT quality, facilitating post-editing effort and boosting translation efficiency.

There exist some limitations to the research. Only one reference translation by a Chinese translator is used, which may not be that neutral in the measurement of automated metrics. And the addition of a second reference translation by an English translator may help in ensuring the reliability of automatic metrics scores. As the study only focuses on word and sentence-level translation quality, future research should continue to explore discourse-level factors such as coherence and style consistency.

References

Castilho, S., Doherty, S., Gaspari, F., & Moorkens, J. (2018). Approaches to Human and Machine Translation Quality Assessment. In *Machine Translation: Technologies and Applications* (pp. 9-38). Springer International Publishing.

Drugan, J., Moorkens, J., Fernández-Parra, M., Rothwell, A., & Austermuehl, F. (2023). *Translation Tools and Technologies*. Routledge.

Fakih, A., Ghassemiazghandi, M., Hafeed Fakih, A., & K. M. Singh, M. (2024). Evaluation of Instagram's Neural Machine Translation for Literary Texts: An MQM-Based Analysis. In *GEMA Online® Journal of Language Studies*, *24*(1), 213-233. Penerbit Universiti Kebangsaan Malaysia (UKM Press).

Wang, H. F., Wu, H., He, Z. J., Huang, L., & Church, K. W. (2022). *Progress in Machine Translation, Engineering*, *18*, 143-153, ISSN 2095-8099. https://doi.org/10.1016/j.eng.2021.03.023

He, L., Ghassemiazghandi, M., & Subramaniam, I. (2024). Comparative assessment of Bing Translator and Youdao Machine Translation Systems in English-to-Chinese literary text translation. In *Forum for Linguistic Studies* (Vol. 6, Issue 2, p. 1189). Bilingual Publishing Group.

Hu, K., & Li, X. (2023). *The creativity and limitations of AI neural machine translation*. Babel. Published online 24 July 2023.

Hutchins, J. (1995). Machine Translation: A brief History. In E. F. K. Koerner & R. E. Asher (Eds.), *Concise history of the language sciences: From the Sumerians to the cognitivists* (pp. 431-445). Pergamon Press.

Hutchins, J. (2007). Machine translation: A concise history. *Mechanical Translation*, *13*(1 & 2), 1-21.

Lavie, A. (2011). Evaluating the Output of Machine Translation Systems. In *Proceedings of the Machine Translation Summit XIII: Tutorial Abstracts, Xiamen, China*, 19-23, September 2011.

Lee, S., Lee, J., Moon, H., Park, C., Seo, J., Eo, S., Koo, S., & Lim, H. (2023). A Survey on Evaluation Metrics for Machine Translation. In *Mathematics* (Vol. 11, Issue 4, p. 1006). MDPI AG.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A. et al. (2022). T*raining language models to follow instructions with human feedback*. arXiv.

Macken, L., Vanroy, B., Desmet, L., & Tezcan, A. (2022). Literary translation as a three-stage process: Machine translation, post-editing and revision. In L. Macken, A. Rufener, J. Van den Bogaert, J. Daems, A. Tezcan, B. Vanroy,…H. Moniz (Eds.), *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation* (pp. 101-110). Ghent, Belgium: European Association for Machine Translation.

Omazić, M., & Šoštarić, B. (2023). New resources and methods in translating legal texts: Machine translation and post-editing of machine-translated legal texts. *Language (s) and Law*, 71-84.

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).

Rivera-Trigueros, I. (2021). Machine translation systems and quality assessment: A systematic review. In *Language Resources and Evaluation* (Vol. 56, Issue 2, pp. 593-619). Springer Science and Business Media LLC.

Rothwell, A., Moorkens, J., Fernández-Parra, M., Drugan, J., & Austermuehl, F. (2023). *Translation Tools and Technologies* (1st ed.). Routledge.

Rothwell, A., Way, A., & Youdale, R. (Eds.). (2023). *Computer-Assisted Literary Translation* (1st ed.). Routledge.

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: "Visions for the future of Machine Translation", Cambridge,* *8-12 August 2006* (pp 223-231).

Thai, K., Karpinska, M., Krishna, K., Ray, B., Inghilleri, M., Wieting, J., & Iyyer, M. (2022). *Exploring Document-Level Literary Machine Translation with Parallel Paragraphs from World Literature* (Version 1). arXiv.

Toral, A., & Way, A. (2015). Machine-assisted translation of literary text. In *Translation Spaces* (Vol. 4, Issue 2, pp. 240-267). John Benjamins Publishing Company.