Morphological and Semantic Evaluation of Ancient Chinese Machine Translation

Kai Jin[1], **Dan Zhao**[1], **Wuying Liu**[2,3]✉
1. School of Foreign Languages, Qilu University of Technology, 250353, Jinan, Shandong, China
2. Shandong Key Laboratory of Language Resources Development and Application, Ludong University, 264025, Yantai, Shandong, China
3. Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, 510420, Guangzhou, Guangdong, China

{jk@qlu.edu.cn, dzhao639@gmail.com, wyliu@ldu.edu.cn}

Abstract

Machine translation (MT) of ancient Chinese texts presents unique challenges due to the complex grammatical structures, cultural nuances, and polysemy of the language. This paper focuses on evaluating the translation quality of different platforms for ancient Chinese texts using *The Analects* as a case study. The evaluation is conducted using the BLEU, LMS, and ESS metrics, and the platforms compared include three machine translation platforms (Baidu Translate, Bing Microsoft Translator, and DeepL), and one language generation model ChatGPT that can engage in translation endeavors. Results show that Baidu performs the best, surpassing the other platforms in all three metrics, while ChatGPT ranks second and demonstrates unique advantages. The translations generated by ChatGPT are deemed highly valuable as references. The study contributes to understanding the challenges of MT for ancient Chinese texts and provides insights for users and researchers in this field. It also highlights the importance of considering specific domain requirements when evaluating MT systems.

1 Introduction

Machine translation (MT) has been a prominent area of research and development in artificial intelligence since the 1950s. Over the years, it has undergone significant advancements, evolving from rule-based methods, statistical methods, and more recently, neural network-based learning methods. As the quality of MT continues to improve and the demand for translation work steadily increases, more and more translators are adopting the “machine translation + post-editing” mode for translation. At the same time, the quality of MT has been a subject of great interest and concern for both the MT and translation fields. Researchers, institutions, and conferences are continuously conducting studies in this area, and various evaluation metrics for MT have been proposed.

There have also been studies on MT of ancient texts. Some researchers have made algorithmic improvements specifically tailored for translating ancient texts (Gutherz et al. 2023; Park et al. 2020; Zhang et al. 2019; Zhou & Liu 2022). Researchers have also conducted evaluations of the quality of MT for ancient texts (Yao et al. 2013; Yang et al. 2021; Yousef et al. 2022). However, research on MT for ancient texts, including ancient Chinese texts, remains relatively scarce.

This paper primarily focuses on MT quality of ancient Chinese texts, and the subsequent discussions will concentrate on this specific domain. Compared to modern Chinese, ancient Chinese has its own unique characteristics. Firstly, ancient Chinese employs complex and distinctive grammatical structures, including syntax, word order, and rhetoric, among other aspects. These structures differ significantly from modern Chinese. MT struggles to accurately capture and parse the intricate grammatical relationships embedded in ancient Chinese texts. Secondly, ancient Chinese texts often employ rhetorical devices such as allusions, symbolism, and metaphors, which involve rich cultural connotations and backgrounds. These allusions and cultural nuances are often challenging for non-Chinese MT systems to comprehend, leading to translation errors or the loss of the original essence and aesthetic appeal. Thirdly, ancient Chinese texts often exhibit polysemy and ambiguity, where a single word or phrase may have multiple interpretations and

Proceedings of the Ancient Language Processing Workshop associated with RANLP-2023, pages 96–102,
held in Varna Bulgaria, Sept 8, 2023.
https://doi.org/10.26615/978-954-452-087-8_011

meanings. MT systems find it challenging to accurately select and judge among these complex semantic relationships, often leading to mistranslations or inaccuracies. The aforementioned characteristics pose significant challenges for MT of ancient Chinese texts.

This study aims to evaluate the translation quality of different platforms for ancient Chinese texts. Through this evaluation, we can gain insights and understanding in dealing with the complexities of ancient language and culture, contribute to the advancement in the field of natural language processing, and provide a supplement to MT quality assessment applications. Furthermore, these evaluation results will help users gain insights into the performance of different platforms, allowing them to identify potential issues and limitations.

2 Experiment design

This study takes the Chinese classic *The Analects*[1] as the research text and compares three classic human-translated versions and four versions generated by four platforms. Three MT quality evaluation metrics are used as evaluation criteria to assess the translation quality of the four platforms. For each human-translated text and each machine-translated text, quality scores are calculated individually. Then, the mean scores are calculated for each platform. The scheme is illustrated in Figure 1.

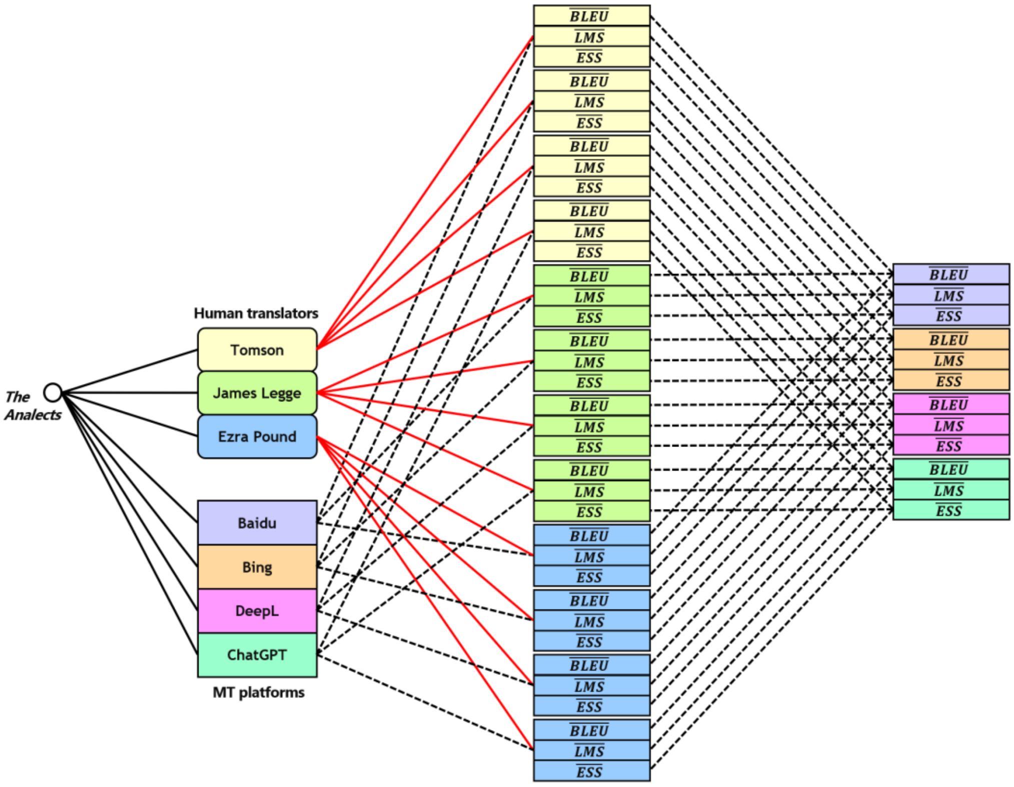

Figure 1: Research scheme.

1

2.1 Texts

In this study, we select *The Analects* as the sample ancient Chinese texts and its three translations as the reference human translations to compare with MT.

The Analects is one of the most influential texts in Chinese ancient philosophy and culture, regarded as a masterpiece in Chinese literature. Its impact extends not only within China but also across the globe, and it has been translated into multiple languages, generating significant influence worldwide (Li & Li 2013). As a

[1] The original Chinese title is “论语” (*lunyu*), and it has several different English versions. In this paper, apart from referring to specific translators, we use “The Analects” to refer to the book.

foundational work of Confucian thought, *The Analects* has garnered the largest number of English translations among classical Chinese texts.

We have compiled the original text of *The Analects* into a corpus, consisting of a total of 1,153 sentences.

We have also selected three highly influential versions of *The Analects* for our study, translated respectively by Tomson (辜鸿铭) (Tomson 2011), James Legge (Legge 2016) and Ezra Pound (Pound 1933). In 1861, James Legge, a missionary from the London Missionary Society, published the first English translation of *The Analects* in Hong Kong. Legge extensively studied the commentaries on *The Analects* from previous generations and used Victorian English in his translation, striving for faithfulness and comprehensiveness. Initially, Legge had a less favorable portrayal of Confucius in his translation. In contrast, Ezra Pound, who identified himself as a Confucian, aimed to transform the world through his translation of Confucian classics (Wang 2004). Pound's translation was published in 1951. Despite his limited proficiency in Chinese, Pound heavily relied on Legge's translation as a reference but also recognized its imperfections, leading him to make significant modifications in his own translation. Pound also emphasized linguistic conciseness (Wei 2005). Another noteworthy translation was by Tomson, published in 1898, which marked the earliest independent Chinese translation of *The Analects*. Tomson had a strong command of multiple languages, a solid linguistic foundation, and extensive knowledge. His English translation of *The Analects* gained wide recognition in the Western world. Tomson believed that Legge's translation often fell short of accurately or fully conveying the original meaning. Thus, Tomson's translation aimed to elucidate the cultural elements missing in the Western context, enabling readers to achieve a more comprehensive understanding (Meng et al. 2012).

In conclusion, while the translations by these three individuals are interconnected, they exhibit distinct characteristics in terms of vocabulary, style, and expression. As highly influential versions, they excel in terms of faithfulness, intelligibility, and elegance in their language. For the aforementioned reasons, we have selected these three translations as reference translations for the purpose of comparing and evaluating machine-translated texts.

2.2 Platforms

The platforms selected for this study include: Baidu Translate ("Baidu" for short), Bing Microsoft Translator ("Bing" for short), DeepL, and ChatGPT[2]. The former three are dedicated online MT systems, while the last one is a conversational generation system based on large-scale language models.

Given that the source text in our research is in ancient Chinese, it is essential for us to select at least one representative MT platforms from China. Baidu is one of the biggest and most influential MT platforms in China. According to industry reports and market data, Baidu consistently ranks first in terms of usage among Chinese MT platforms[3]. Therefore, Baidu has become our top choice as a MT platform developed in China.

For MT platforms outside of China, we have chosen Bing and DeepL. Both platforms are widely recognized and highly regarded for their usage and performance worldwide. Based on our extensive translation practice, we have observed that DeepL's translations occasionally exhibit noticeable differences in vocabulary and even sentence structure compared to other MT platforms.

ChatGPT is a language generation model that possesses the capability to comprehend and produce natural language text, encompassing translation tasks as well. While its primary utility lies in dialog and text generation, it can, to a certain extent, engage in translation endeavors. This attribute permits viable comparisons with conventional MT systems under specific circumstances. Recently, ChatGPT's performance in translation tasks has gained increasing attention and recognition. Although there is currently limited research on the translation quality of ChatGPT, some researchers have already drawn the conclusion that "ChatGPT has already become a good translator." (Jiao et al. 2023) Based on our observation, we have also found that the translations generated by ChatGPT exhibit differences from the three MT platforms. It is worth noting that each generation of translation by ChatGPT can vary, and the translation can also be adjusted based on the given prompts. Therefore, to

[2] ChatGPT-3.5 version is utilized in this study.

[3] Information source: http://bjx.iimedia.cn/app_rank, last accessed 2023/6/27.

ensure relatively reliable experimental results, we only select the first-generation translation produced by ChatGPT without adding any other prompts than “Translate… into English.”

3 Evaluation metrics

The evaluation metrics adopted in this research include Bilingual Evaluation Understudy (BLEU) (Papineni et al. 2002), Levenshtein-distance-based Morphological Similarity (LMS) and Pretrained-model-based Embedding Semantic Similarity (ESS).

3.1 BLEU

In 2002, IBM proposed the BLEU metric, which has become the de facto standard for evaluating MT quality. This metric is based on the mechanical morphological evaluation method using n-gram grammar. In this paper, the BLEU referred to is BLEU4.

$$\overline{BLEU} = \frac{1}{nm}\sum_{j=1}^{n}\sum_{i=1}^{m} BLEU(r_{ij}, c_i) \quad (1)$$

For a specific application scenario involving a machine-translated text collection (***C***) consisting of *m* sentences and the corresponding *n* sets of human reference translations (***R***), we evaluate using the arithmetic mean $\overline{BLEU}$, as shown in equation (1), of the BLEU metric.

3.2 LMS

To evaluate the morphological similarity between sentences, we introduce the LMS metric. This metric is based on the edit distance proposed by the Soviet mathematician Vladimir Levenshtein in 1965. The edit distance refers to the minimum number of editing operations required to transform one string into another, including substitution, insertion, and deletion. Let *LD*(r, c) represent the edit distance between a human reference translation (*r*) and a machine-translated candidate (*c*). The equation (2) represents the LMS. In the equation, length (*r*) represents the length of the reference translation and length (*c*) represents the length of the candidate translation. The LMS value ranges from 0 to 1, where a higher value indicates a greater morphological similarity between the sentences.

$$LMS(r, c) = 1 - \frac{LD(r, c)}{Max(Len(r), Len(c))} \quad (2)$$

For a specific application scenario involving a machine-translated text collection (***C***) consisting of *m* sentences and the corresponding *n* sets of human reference translations (***R***), we evaluate using the arithmetic mean $\overline{LMS}$, as shown in equation (3), of the LMS metric. In this experiment, the getLevenshteinDistance library function from org.apache.commons.lang3.StringUtils is used.

$$\overline{LMS} = \frac{1}{nm}\sum_{j=1}^{n}\sum_{i=1}^{m} LMS(r_{ij}, c_i) \quad (3)$$

3.3 ESS

To address the challenge of handling synonymous and morphologically variant expressions, we introduce the ESS metric as a semantic similarity evaluation index. This metric maps the human reference translation (*r*) and machine-translated candidate (*c*) to embedding vectors in a pre-trained model (Peters at al. 2018). Specifically, we obtain the embedding vectors ($\boldsymbol{v_r}$) for the reference translation and ($\boldsymbol{v_c}$) for the candidate translation. Then, we calculate the cosine similarity between vectors $\boldsymbol{v_r}$ and $\boldsymbol{v_c}$ in the embedding vector space (Reimers & Gurevych, 2019), representing the embedding semantic similarity between the reference and candidate translations as *ESS(r, c)*. According to the definition of this metric, the embedding semantic similarity values between two sentences is within [-1, 1]. To further normalize these values so that *ESS(r, c)*∈[0, 1], we apply a proportional scaling transformation.

For a specific application scenario involving a machine-translated text collection (***C***) consisting of *m* sentences and the corresponding *n* sets of human reference translations (***R***), we evaluate using the arithmetic mean $\overline{ESS}$, as shown in equation (4), of the ESS metric. In this study, the all-roberta-large-v1 pre-trained model[4] was used.

[4] https://huggingface.co/sentence-transformers/all-roberta-large-v1

$$\overline{ESS} = \frac{1}{nm}\sum_{j=1}^{n}\sum_{i=1}^{m} ESS(r_{ij}, c_i) \quad (4)$$

4 Experiment results and analysis

First, we compare each human-translated text with each machine-translated text respectively under the three metrics BLEU, LMS and ESS, and the results are shown in Table 1. The highest value obtained when comparing the texts from different platforms to the same human translator is highlighted in bold. We can see that, except for one LMS value from DeepL, all the other highest values belong to Baidu.

Platforms	Human translators	Metrics		
		$\overline{BLEU}$	$\overline{LMS}$	$\overline{ESS}$
Baidu	Tomson	**0.1059**	0.2857	**0.8494**
	James Legge	**0.4901**	**0.3731**	**0.9162**
	Ezra Pound	**0.539**	**0.5382**	**0.9516**
Bing	Tomson	0.0469	0.2987	0.8109
	James Legge	0.0251	0.239	0.8468
	Ezra Pound	0.0905	0.3868	0.8597
DeepL	Tomson	0.0621	**0.3323**	0.8408
	James Legge	0.0356	0.2656	0.8611
	Ezra Pound	0.1049	0.3731	0.8321
ChatGPT	Tomson	0.0474	0.2996	0.8117
	James Legge	0.0253	0.2408	0.8478
	Ezra Pound	0.0907	0.3878	0.8606

Table 1. BLEU, LMS and ESS results of human-translated texts and machine-translated texts.

Then, we calculate the mean of the three values for each platform under each metric, resulting in the evaluation results for the translation quality of each platform. The results are shown in Table 2. It can be observed that Baidu has the best performance under all the metrics, with the BLEU value significantly surpassing the other three platforms.

Platforms	$\overline{BLEU}$	$\overline{LMS}$	$\overline{ESS}$
Baidu	**0.3783**	**0.3990**	**0.9057**
Bing	0.0542	0.3078	0.8391
DeepL	0.0545	0.3091	0.8400
ChatGPT	0.0675	0.3234	0.8446

Table 2. Evaluation results of the four platforms.

In this study, Baidu outperforms other platforms in translating ancient Chinese texts into English. Aside from technical architecture and optimized algorithms, the most possible reason is that Baidu, as a Chinese company, may have access to a larger and more diverse dataset of bilingual English-Chinese texts compared to other platforms, allowing their models to learn from a wide range of language patterns and translation examples. This extensive training data contributes to improving the accuracy and fluency of it Chinese-English translations.

5 Other findings: ChatGPT's advantages

From Table 2 we can observe that ChatGPT has the second-best performance in translation, right after Baidu. Also, we have found that ChatGPT has also demonstrated exceptional advantages in certain aspects.

(1) Semantic interpretation

In ancient Chinese, the phenomenon of polysemy, where a single word has multiple meanings, is very common. MT must make accurate judgments regarding the selection of the intended meaning for such words. These judgments often rely on the contextual information provided by the surrounding text. However, when the context does not provide sufficient information, MT systems often make errors in determining the correct meaning. On the other hand, ChatGPT demonstrates superior performance in the disambiguation of polysemous words, excelling in making accurate sense selection judgments even in cases where context is limited.

Example 1

Source Text ("ST" for short): 君子不器。

Baidu: A gentleman is not a weapon.

Bing: A gentleman is not a weapon.

DeepL: A gentleman is not a tool.

ChatGPT: A noble person is not defined by their possessions or material objects.

In this context, “器” refers to “utensil.” This sentence conveys the idea that a noble person (君子) should have a broad and inclusive mindset, rather than being limited to a specific role or function like a utensil (器具). However, Baidu and Bing mistakenly interpret “器” as “weapon” (武器) and DeepL “tool” (工具).

Example 2

ST: 诲女知之乎！

Baidu: Teach women to know!

Bing: The girl knows!

DeepL: Teach your daughter to know!

ChatGPT: Have you learned to understand this?

In ancient Chinese, the character “女” (female) is often used as a phonetic loan character, representing the pronoun “汝” (thou). In Example 2, Baidu, Bing, and DeepL do not correctly identify the true intended meaning of “女,” while ChatGPT do. We have found that, in our research, the translation of all instances of the character “女” as a phonetic loan character in *The Analects* exhibits the same pattern across different MT translations.

(2) Fault tolerance

During the process of data compilation, we have discovered that there were several errors in the original text (These errors were meticulously rectified, ensuring the source text’s accuracy prior to commencing the MT task). When the text with errors is input into the platforms, Baidu, Bing and DeepL generate translations based on the faulty text. However, ChatGPT often demonstrate the ability to recognize and identify these errors and generate translations based on the correct characters. ChatGPT possesses this capability that is not present in traditional MT systems.

Example 3

ST: 邦君之妻，君称之曰夫人，夫人自称日小童……

Baidu: Bangjun’s wife, whom you call Madame, and whose wife calls herself Rixiaotong…

Bing: The wife of the prince, the prince is called Mrs. Wang, and the lady calls herself Ri Xiaotong…

DeepL: The wife of the king of the state, the king called it the wife, the wife called herself the day child…

ChatGPT: The wife of the ruler of a state is called “Madam” by the ruler, and she refers to herself as “Little Child”…

In Example 3, the underlined character “日” (pronounced as “*ri*”, meaning the sun or day) in the source text should be “曰” (say), and “小童” that follows means “little child”. This sentence means that the madam calls herself “小童.” Both Baidu and Bing transliterate this character mistakenly and DeepL interprets “日” as “day,” while ChatGPT recognizes this fault.

Example 4

ST: 恭而无礼则劳，慎而无礼则蒽……

Baidu: Respectful but impolite leads to labor, cautious but impolite leads to anthracene…

Bing: If you are respectful and rude, you will work, if you are cautious and rude, you will be rude…

DeepL: Respectful but no ritual is labor, prudent but no ritual is anthracite…

ChatGPT: If one is respectful without propriety, it leads to laboriousness. If one is cautious without propriety, it leads to timidity…

“蒽” (anthraquinone) in the original text is a faulty character which should be “葸” (afraid, bashful). Both Baidu and DeepL take “蒽” directly, and Bing simply omits it. ChatGPT recognizes the fault and translates it into “timidity,” which aligns with the intended meaning of the original text.

6 Conclusion

Based on the evaluation of translation quality of the Chinese classic *The Analects* by Baidu, Bing, DeepL, and ChatGPT using the BLEU, LMS, and ESS metrics, we have found that among the four platforms, Baidu, as a MT platform developed in China, performs the best in handling ancient Chinese texts. Its scores in all three metrics are significantly higher than the other three platforms. ChatGPT, as a general-purpose language model, ranks second among the four, and has demonstrated unique advantages, and the translations it produces are highly valuable as references. It is worth mentioning that in this study, the translations generated by ChatGPT were done without any prompts (except for the one mentioned in 2.2) or adjustments. We plan to discuss in our future research the translation quality of ChatGPT by incorporating prompts for adjusting the translation of ancient Chinese into English.

Acknowledgments

The research is supported by the Scientific Research Project of China National Committee for Terminology in Science and Technology (No. YB2021027), the Art Science Key Project of Shandong Provincial Association for Science of Arts & Culture (L2022Z06170304), the New Liberal Arts Research and Reform Practice Project of Ministry of Education of China (No. 2021060049), the Postgraduate Education and Teaching Reform Research Project of Shandong (No. SDYJG21185), the Key Project of Undergraduate Teaching Reform Research of Shandong (No. Z2021323), the Science and Technology Program of Guangzhou (No. 202201010061), and the Humanity and Social Science Research Project of Ministry of Education of China (No. 20YJAZH069).

References

Kishore Papineni, Salim Roukos, Todd Ward, et al. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311-318.

Gai Gutherz, Shai Gordin, Luis Sáenz, et al. 2023. Translating Akkadian to English With Neural Machine Translation. *PNAS nexus*, 2(5): pgad096.

Chanjun Park, Chanhee Lee, Yeongwook Yang, et al. 2020. Ancient Korean Neural Machine Translation. *IEEE Acces*, 8: 116617-116625.

Zhiyuan Zhang, Wei Li, Qi Su. 2019. Automatic Translating Between Ancient Chinese and Contemporary Chinese with Limited Aligned Corpora. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*. Springer International Publishing, 157-167.

Chengbin Zhou, Zhongbao Liu. 2022. Machine Translation of Ancient Chinese Text Based on Transformer of Semantic Information Sharing. *Technology Intelligence Engineering*, 8(06): 114-127.

Zhenjun Yao, Xuhong Zheng, Pengtao Xu, et al. 2013. An Exploration of Phrase-Based SMT for English Translation of Tao Te Ching. *Shandong Foreign Language Teaching*, 34(03): 109-112.

Kexin Yang, Dayiheng Liu, Qian Qu, et al. 2021. An Automatic Evaluation Metric for Ancient-Modern Chinese Translation. *Neural Computing and Applications*, 33: 3855-3867.

Tariq Yousef, Chiara Palladino, David J. Wright, et al. 2022. Automatic translation alignment for ancient Greek and Latin. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, 101-107.

Gang Li, Jinshu Li. 2013. On the English Translation of *The Analects*: A Survey. *Journal of Social Science of Hunan Normal University*, 42(01).

Tomson. (Trans.). 2011. *The Discourse and Sayings of Confucius*. Yunnan People’s Publishing House, Kunming, China.

James Legge. (Trans.). 2016. *Confucian Analects*. Liaoning People’s Publishing House, Shenyang, China.

Ezra Pound. (Trans.). 1933. *Confucian Analects*. Peter Owen Limited, London, UK.

Hui Wang. 2004. A Comparison of Confucian Analects Translated by James Legge and Ezra Pound. *Foreign Language and Literature*, (05), 140-144.

Wangdong Wei. 2005. A Multi-Perspective Comparison of Three Translations of Lun Yu: From James Legge, Ezra Pound to Edward Slingerland. *Chinese Translators Journal*, (03), 52-57.

Jian Meng, Tao Qu, Yang Xia. 2012. English Translation of Chinese Classics under Adaptation Theory – Reflections on the English Translation of *Lunyu* by Gu Hong-ming. *Foreign Language Research*, (03), 104-108.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, et al. 2023. Is ChatGPT a Good Translator? A Preliminary Study. *arXiv preprint arXiv:2301.08745*.

Kishore Papineni, Salim Roukos, Todd Ward, et al. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311-318.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, et al. 2018. Deep Contextualized Word Repretations. *arXiv:1802.05365*.

Nils Reimers, Iryna Gurevych. 2019. Sentence-Bert: Sentence Embeddings Using Siamese Bert-Networks. *arXiv preprint arXiv:1908.10084*.