Literature, Language and Cultural Studies
ISSN Print: 3079-5095
ISSN Online: 3079-5109
https://doi.org/10.63313/LLCS.9057

Cross-Language Translation Evaluation: A Comparative Analysis of the Quality of English-Chinese Interpretation by AI Models and Strategies for Improvement

Tong Su
Hangzhou Normal University of Foreign Languages, Hangzhou 311121, China
Email: xlin02@126.com

How to cite this paper: Su, T. (2025). Cross-Language Translation Evaluation: A Comparative Analysis of the Quality of English-Chinese Interpretation by AI Models and Strategies for Improvement. Literature, Language and Cultural Studies, 2(1), 32-40. ISSN Print: 3079-5095; ISSN Online: 3079-5109.
https://doi.org/10.63313/LLCS.9057
Published: 2025-07-03

Abstract

Machine translation, as one of the important functions of artificial intelligence, plays an increasingly important role in cross-linguistic communication. The paper is meant to evaluate the performance of domestic up-and-coming AI models in English-Chinese mutual translation tasks in order to analyze their functions and effects comparatively and explore strategies to improve the quality of AI translation in the process of both 2 languages. Based on the literature review in the first section, we put forward the hypothesis that there are certain deficiencies in AI translation tasks, and that different types of errors will be produced depending on the target language. We will use both quantitative and qualitative analysis methods, combined with example analysis and the application of the BLEU model, to evaluate the AI translation output in terms of several dimensions. We also point out the limitations of the study and discuss the main problems of AI translation and propose targeted enhancement strategies in the conclusion part.

Keywords

AI Models; English-Chinese Translation; BLEU Model; Translation Quality; English-Chinese Comparison

1. Introduction

AI translation mimics the cognitive computational features of the human brain in the translation process, thus speed is one of its major advantages compared with human translation. However, the quality of AI translation has always been the most fatal problem due to the profoundness of language. Current studies on AI translation mainly focus on the functions of AI translation or compare human-machine translation from different levels, revealing the characteristics of machine translation that are different from those of human translation, but there are also some shortcomings: almost all of them concentrate on AI translation itself or just compare machine

translation and human translation, but few of them study cross-linguistic aspects. Nowadays, there are many up-and-coming AI models in China, such as WenxinYiyin and Kimi, etc. When dealing with English-Chinese translation, will domestic AI perform differently in English-Chinese or Chinese-English translation? Studying this phenomenon and the root causes of this problem is meaningful for the quality improvement of AI translation and that is the purpose of this study.

2. Hypothesis

Currently, there is a lack of research on AI translation quality based on the English-Chinese comparison perspective. Will foreign AI models such as ChatGPT perform differently in Chinese-to-English and English-to-Chinese translation compared to domestic AI models?

Based on the literature review and the identified gaps in cross-language AI translation research, this study proposes the following hypotheses:

Hypothesis1: There will be a significant difference in translation quality between English to Chinese and Chinese to English when using domestic AI models, with one direction showing higher accuracy and fluency.

Hypothesis2: The type and frequency of translation errors in ETC and CTE tasks will differ due to language and cultural differences, with certain error categories being more prevalent in one direction.

Hypothesis3: Different AI intelligence models will perform differently when dealing with translation tasks, depending on the ways to train AI models.

These hypotheses will be tested using a combination of quantitative and qualitative analysis methods, including BLEU evaluation and specific evaluation of translation samples. The results of the study will help provide insights into the performance of domestic AI models in English-Chinese translation tasks and inform strategies for improvement.

3. Methodology

This section is an introduction to the methodology of this study. We will select a number of articles that have a relatively standard reference translation of each other, and then ask the domestic AI models Wenxinyiyin, Kimi and KU Xunfei to translate the given articles from Chinese to English and from English to Chinese. Then record the BLEU scores, which will be shown in a table below so that we can understand the scores intuitively.

After BLEU test, this study will also select one of the translations for a more specific analysis of the results of the inter-translation, including judgement on word choices, sentence patterns, etc. The purpose is to detect whether the problems that occur when the AI intelligent model handles the task of translating from English to Chinese and whether they are of the same types, and if not, then what are the aspects that trigger off the problems, and whether they are the same as those of the human

translators' difficulties? Are they consistent?

Before starting the research, this paper will give a brief introduction to the BLEU mechanism, although no specific algorithms will be presented in this paper. Subsequently, in the second part of this section is an example analysis.

3.1. BLEU models

BLEU stands for Bilingual Evaluation Understudy, which is an algorithm commonly used to evaluate the quality of machine translation, proposed by IBM. It evaluates the accuracy of translation by comparing the n-gram overlap between the automatic translation results and one or more reference translations.

The BLEU algorithm consists of four steps:

1. N-gram calculation: n-grams are calculated between the translation results and the reference translations to get the number of n-gram overlaps between them.

2. Exact Matching: For each n-gram, calculate the smaller values of the number of occurrences in the automatic translation result and the number of occurrences in the reference translation, and then add these smaller values to get the total number of exact matches.

3. Brevity Penalty: if the length of the automatic translation is less than the length of the reference translation, the BP penalty is 1, otherwise it is exp, that is, 1 minus the length of reference translation or the length of automatic translation.

4. Score Calculation: Divide the number of exact matches by the total number of n-grams of the automatic translation result, and then multiply it by the BP penalty to finally get the BLEU score.

Table 1. Result of the BLEU model.

BLEU Scores				
	TEXT1	TEXT2	TEXT3	TEXT4
Kimi	0.68	0.75	0.61	0.76
Wenxinyiyan	0.71	0.79	0.66	0.71
KU Xunfei	0.81	0.66	0.78	0.86
Kimi	0.54	0.66	0.59	0.33
Wenxinyiyan	0.47	0.57	0.76	0.51
KU Xunfei	0.64	0.61	0.81	0.63

In this paper, four original manuscripts in Chinese and English were selected and then we asked the three AI models to translate the Chinese version into English, and the English version into Chinese. Here are the corresponding BLEU scores (between 0 and 1, with two decimal places):

It can be seen from the table that Ku Xunfei showed high translation quality in most cases, especially in Chinese-to-English translations. However, Kimi performs better on some texts in the English to Chinese translations, while some are even worst if not given certain order to ask it to refer to relevant data for the purpose of improv-

ing the translation quality. Wenxinyiyan performs relatively consistently in the Chinese to English translations, but scores lower in the English to Chinese translations. TEXT4 is a legal text, and none of the three AI models performed very well in the English to Chinese translation session, while TEXT3 is a literary text, and the latter two scored higher relative to Kimi when there is big data available for reference. Finally, the three AI models perform better in handling CTE translation.

3.2. Detailed analysis

Here we also chose additional Constitutional documents material for human analysis to test whether the BLEU meets our expectations on the one hand, and on the other hand to be able to explore the reasons for the test results in relation to social factors such as cultural context.

Chinese-English

P1：

Reference:All ethnic groups of the People's Republic of China are equal.

KU Xunfei: All nationalities in the People's Republic of China are equal.

Nationality is understood internationally as “country of belonging” as “Nation” and “ethnic group” are two very different concepts in foreign literature. They represent totally different meanings, and express different forms of identity of human societies in different historical scenarios. Today, the official website of the State People’s Committee uses the term “ethnic groups”.

P2:

Reference：The state shall protect the lawful rights and interests of all ethnic minorities and uphold and promote relations of equality, unity, mutual assistance and harmony among all ethnic groups.

Kimi: The state protects the legitimate rights and interests of all ethnic minorities and maintains and develops equal, united, and mutually supportive relationships among all ethnic groups.

In legal texts, “shall” is usually used to express a mandatory obligation, so the original translation is more accurate on this point. In legal texts, “lawful” more accurately conveys the meaning of being protected by law. The word “legal” is also used here.

Secondly, supportive relationships omits the meaning of equality, solidarity and mutual assistance, which as a translation of legal documents needs to be rigorous and not lacking in quality.

P3:

Reference: Discrimination against and oppression of any ethnic group are prohibited; any act that undermines……

Kimi：Discrimination and oppression against any ethnic group are prohibited, as are……

Wenxinyiyan: Discrimination and oppression against any ethnic group are prohibited, as well as……

Semicolons provide a clear separation while maintaining sentence coherence. In legal texts, where clarity and accuracy are paramount, semicolons can help readers better understand sentence structure.

P4:

Reference：The state shall, in light of the characteristics and needs of all ethnic minorities, assist……

KU xunfei：The state assists areas inhabited by minority nationalities in accelerating their economic and cultural development according to the characteristics and needs of the various minority nationalities.

Here the reference uses “in light of” to bring forward the second half of the sentence, emphasizing that China takes into account the characteristics and needs of ethnic minorities when taking action. This suggests that the policy or action is not arbitrary, but is based on careful consideration of the situation of the particular group.

English-Chinese

P5

Reference：All ethnic groups of the People's Republic of China are equal. (中华人民共和国各民族一律平等。)

Kimi: 中华人民共和国的所有民族都是平等的。

KU xunfei: 中华人民共和国的所有民族都享有平等权利。

“Yi lv” carries a certain degree of formality and authority in Chinese, and is often used in official documents such as laws and policies to emphasize the universal applicability and inviolability of rules or principles. When translating texts involving national policies or legal statements, the use of “Yi lv” can better convey the seriousness and authority of the original text.

P6:

Reference：The state shall, in light of the characteristics and needs of all ethnic minorities, assist all ethnic minority areas in accelerating their economic and cultural development. (国家根据各少数民族的特点和需要，帮助各少数民族地区加速经济和文化的发展。)

Kimi：国家将根据所有少数民族的特点和需要，帮助所有少数民族地区加快其经济和文化发展。

KU xunfei：根据各少数民族的特点和需求，国家将帮助少数民族地区加快发展经济和文化。

KU xunfei has made some adjustments to the order of the words in the translation, placing the phrase “according to the characteristics and needs of the various ethnic minorities” at the beginning of the sentence, which makes the basis for the State’s action more prominent at the beginning of the sentence, but it does not conform to the custom of the Chinese language, where the subject is seldom placed in the middle of the sentence.

Literature Texts

P7:

Reference:

Frailty, thy name is woman!

脆弱啊，你的名字是女人!

Kimi:

脆弱，你的名字是女人

"O fragility, thy name is woman!"

Wenxinyiyan:

脆弱啊，你的名字就是女人！

Frailty, thy name is woman!

This is a well-known segment of Hamlet, here the processed content of KU xunfei is no different from the reference translation and we are aware of the fact that one of the advantages of AI is it can invoke big data, as it can find relevant literature and comments written by scholars and then translated. Kimi is not doing well, obviously it does not make use of the big data but directly translates the given text. Although the meaning is right, some content is redundant compared to the reference translation.

P8:

Reference: The time is out of joint – O, cursed spite, that ever I was born to set it right!

这是一个礼崩乐坏的时代，唉!倒霉的我却要负起重整乾坤的责任。

Kimi:

时间已经错乱——哦，该死的遗憾，我竟然生来就是为了纠正它！

KU xunfei:

这个时代已经混乱不堪——哦，该死的厄运，我竟然出生来纠正它！"

It can be found that it is difficult for the AI model to handle the English translation when it does not call on big data, and not only is it weaker in terms of tone and expression of emotion, but also the wording is not very accurate, compared to the scholars' version used as the reference translation. Translation is a complex task involving linguistic accuracy, understanding of cultural context and transmission of emotions. The AI translation model may not be able to reach the level of a professional translator without sufficient contextual information and in-depth understanding of the original text's emotion and context. In addition, AI translators may lack flexibility and creativity to a certain degree.

4. Discussion and Limitation

4.1. Discussion

Through the BLEU test and human analysis, this paper can also summarize certain laws.

First of all, BLEU conducts a generalized test on the translation results of the three AI models, and we can find that these models perform better in dealing with the

tasks of Chinese-to-English translation than English-to-English translation. This question seems to be answered by some elements in the detail analysis part. On the one hand, the English to Chinese translation needs to take into account the correspondence of terminology, and Chinese may have more choices in vocabulary expression compared to English. On the other hand, the AI model is less flexible in sentence changes, that is, it is not in line with the Chinese expression habits. In comparison, the domestic AI translation models show higher translation quality when dealing with the Chinese-to-English translation task. This may be due to the fact that in the CTE translation task, the AI can use its big data analysis capability to find more accurate English counterparts and expressions, but regarding the ETC translation, there are fewer information resources in the database. One of the advantages of AI translation is that it can use big data resources to find translations that are more in line with the context of the original text. However, when the AI model fails to utilize this advantage effectively, the quality of translation may be worse.

In addition, in the task of translating some specific texts, some AI performs particularly well as it finds a widely circulated and popularly recognized version and presents it directly to the user, such as legal texts. The biggest advantage of AI translation over traditional machine translation is the ability to "find the context that the machine cannot find". For example, when translating certain types of text, such as legal texts, contracts, and policy documents, AI can find widely circulated and universally accepted translations. These versions are usually reviewed and verified by professional translators and therefore have a high degree of accuracy and authority. At the same time, AI translation systems, through continuous learning and optimization, are able to better understand and handle the complexity and professionalism of a particular text. With the continuous enrichment of data resources and the improvement of algorithms, the translation ability of AI in these fields is expected to be further improved.

Of course, there are deeper methodological modifications, but they require interdisciplinary techniques. For example, by utilizing the information in multi-layer encoders and decoders, different levels of syntactic and semantic information can be captured, thereby improving the accuracy and fluency of translation. For example, the method proposed by Tencent AI Lab to improve the Transformer model through layer aggregation and multi-layer attention model, simultaneously utilizing information from different network layers.

In conclusion, although the domestic AI translation model performs better in the CTE task, it still needs to be improved in the ETC task, especially in terms of terminology correspondence, sentence flexibility, cultural-emotional transfer, and the use of big data resources. Through continuous optimization and learning, the AI translation model is expected to reach a higher level of translation quality in the future.

4.2. Limitation

Previous study is all about research itself and discussion, but of course there are certain drawbacks as well as incompleteness in this article. The main points are summarized as follows:

The first question is sample selection bias. As we know, the limited number of samples used in the study or the unrepresentative sample selection may affect the generality and reliability of the research results. The articles used for example analysis in this article include legal documents, literary works, news reports, etc., which are mainly selected as some text types with wider audiences, and the main condition is that there are corresponding reference translations in order to test the validity of the BLEU mechanism. Therefore, the number of samples is still too small, so this paper can be regarded as a reference in the example analysis part. For a more in-depth study, more comprehensive and extensive testing is still needed.

The second one is the limitation of the BLEU model. Although the BLEU model is widely used to evaluate the quality of machine translation, it mainly focuses on the similarity of the surface form, which may not be sufficient to comprehensively evaluate the deep semantics and contextual adaptation of translation. In addition to the BLEU model, there are some other models that can measure the quality of translation, such as LQA, DQF and so on.

The third one is the problem of the application area of the research results. This article discusses the limitations and improvement strategies of the current domestic AI models in dealing with translation tasks, but it does not explain in detail how to apply these research results to the actual AI translation model improvement, and it is not quite clear how to train the AI to improve the performance of the AI in translation tasks after knowing the direction of improvement.

The fourth one is that the specificity of the research hypotheses could be improved. The three hypotheses proposed in this paper, while providing direction for the research, may need to be more specifically defined and operationalized in order to be tested in empirical studies.

5. Conclusion

This study aims to evaluate the performance of domestic AI models in English-Chinese translation tasks, and assesses the quality of AI translation output in multiple dimensions through quantitative and qualitative analysis methods, combined with the technology of BLEU model.

In this study, some articles with relatively standard reference translations were selected, and then the domestic AI models Wenxinyiyin, Kimi, and KU Xunfei were allowed to translate them, as BLEU scores were recorded. In addition, this study will analyze one of the translations more specifically, including judgments on vocabulary selection, sentence structure and so on, in order to detect the types of problems that AI intelligent models have when dealing with the English to Chinese translation task and to explore the causes of the problems. Through BLEU and further research, it

has been shown that there are differences in the performance of AI translation models in English-Chinese translation tasks, especially in the translation of some specific texts. Accuracy of AI translation, understanding of cultural context, and transfer of emotions are key factors in translation quality. Despite their speed advantages, AI translators may not be able to reach the level of professional translators when they lack sufficient contextual information and in-depth understanding of the emotions and background of the original text.

Based on the above contents, we believe that future research can further explore the effectiveness of AI translation in specific language environments, as well as the cross-cultural factors that contribute to these differences, thus making it possible to further improve the way in which AI models are trained so that these AI models perform equally well in inter-translation tasks.

References

[1] Briva-lglesias,V., B.R, Cowan &,S. O' Brien. The impact of traditional and interactive post-editing on Machine Translation User Experience, quality, and productivity [J]. Translation,Cognition&Behavior,2023(6): 58-84.

[2] Koglin, A. An empirical investigation of cognitive effort required to post-edit machine translated metaphors compared to the translation of metaphors[J]. Translation & Interpreting, 2015(1):126-141.

[3] Laskar, M,T.R.,et al. 2023. A Systematic study and comprehensive evaluation of ChatGPTon Benchmark Datasets[J]. arXiv preprint arXiv, 2023.

[4] Liu,P.F.,et al. Pre-train, prompt, and predict:A systematic survey of promptingmethods in natural language processing[J]. ACM Computing Surveys, 2023(9): 1- 35.

[5] Jiejing, P. Research on the Quality Evaluation Model of Corpus-Assisted Intertextual Translation of Business Texts [J] Journal of Mudanjiang Institute of Education, 2023(10): 21-29.

[6] Chunhua, S. Optimization of English Machine Translation Based on Artificial Intelligence Deep Neural Network [J]. Modern Electronics Technology, 2024, 47(03): 80-84.

[7] Yiwei, Y. A Comparative Study on the Quality of Human-Machine English Translation [D]. Shanghai International Studies University, 2023.

[8] Lu, W, Xiangling, W. Research on the Training Mode of Post-Editing Ability of Machine Translation in the ChatGPT Era [J]. Foreign Language Audio-Visual Teaching, 2023(4):16-23+115.