English Language Teaching and Linguistics Studies
ISSN 2640-9836 (Print) ISSN 2640-9844 (Online)
Vol. 5, No. 2, 2023
www.scholink.org/ojs/index.php/eltls

*Original Paper*

# Quality Evaluation of C-E Translation of Legal Texts by Mainstream Machine Translation Systems—An Example of DeepL and Metasota

Ashley Yu[1]

[1] ECUPL, Shanghai, China

Received: April 6, 2023 Accepted: May 16, 2023 Online Published: May 23, 2023

doi:10.22158/eltls.v5n2p180 URL: http://dx.doi.org/10.22158/eltls.v5n2p180

***Abstract***

*Despite significant progress made in machine translation technology and the ongoing efforts in practical and commercial application of neural machine translation systems, their performance in vertical fields remains unsatisfactory. To avoid misunderstandings and excessive expectations of a specific machine translation system, this research selected legal texts as its real data research object. The text translation tasks were accomplished using two popular neural machine translation systems, DeepL and Metasota, both domestically and internationally, and evaluated using internationally recognized BLEU algorithm to reflect their Chinese-to-English translation performance in legal fields. Based on the determined BLEU score, the study adopted an artificial analysis method to analyze the grammatical aspects of the machine translation output, including the accuracy of terminology usage, word order, subject-verb agreement, sentence structure, tense, and voice to enable readers to have a rational understanding of the gap between machine translation and human translation in legal text translation, and objectively assess the application and future development prospects of machine translation in legal text fields. The experimental results indicate that machine translation systems still face challenges in achieving high-quality legal text translations and meeting practical needs, and that further post-translation editing research is needed to improve the accuracy of legal text translation.*

***Keywords***

*neural machine translation, translation evaluation, legal text, comparative analysis*

## 1. Introduction

### *1.1 Research Background*

With the development of economy and the progress of our society, new breakthroughs have been made in the technology of machine translation. Nowadays, machine translation has a more and more complex and perfect grammar process program. The statistical machine translation technology has been initially applied to the machine translation system. From the perspective of the application of machine translation, machine translation has been applied in daily life more and more frequently in recent years. Currently, there are a large number of online machine translation systems that provide near-instantaneous translations in almost all fields, far faster than human translations, and their translation quality varies from system to system, from language to language, and from text to text.

However, whether the translation quality of translation can meet the needs of people, especially whether the translation quality of specialized machine translation can meet the requirements of the specified industry still remains under explored.

Long and complex sentences, irregular word order, combined with the specialized terminology of the legal systems used, in a sense make legal texts a special type of language, so machine translation quality assessment is necessary to clarify which online machine translation system is the best choice for Chinese to English translation of legal texts.

Machine translation assessment has gained traction in the more than six decades since the first ALPAC report (Note 1). The industry is usually divided into two types of assessments: manual assessments and automated machine assessments. Manual assessments, however, inevitably include inconsistent judgments due to personal biases and preferences, resulting in unreliable evaluations. Manual evaluation also has many drawbacks, such as time-consuming, expensive, irreducible, and in many cases inconsistency (subjectivity) among manual evaluators. Therefore, automated evaluation methods have become both a technical and practical necessity.

The purpose of this paper is to evaluate the translation quality of two popular machine translation systems in the field of legal translation from Chinese to English, and to identify which online machine translation system is the best for automatic translation of legal texts from Chinese to English, using the advanced BLEU quantitative machine translation evaluation technique to compare them horizontally. choice. This technique has been generally accepted and widely used by machine translation developers and researchers in recent years, and can undoubtedly provide a more reliable objective rather than subjective understanding of the quality of legal texts translated by machine translation systems.

In the following sections, the author first provides a brief review of the relevant literature on the topic and introduces the BLEU assessment instrument, followed by a description of the methodology, including the selection of test texts and the assessment process, and finally summarizes the main findings, limitations of the study and provides suggestions for future research.

*1.2 Significance of the Research*

The significance of this study lies in its ability to provide a better comprehension of the performance of machine translation systems in translating Chinese legal text into English. In addition, it identifies the optimal online machine translation system for automatically translating legal texts from Chinese to English out of the two systems tested. The study utilized the BLEU evaluation method to conduct a quality assessment of the translation outputs produced by leading machine translation systems in the Chinese-English translation field, DeepL and Metasota. Such assessments are vital for clients seeking to determine the optimal choice of machine translation system for their specific requirements.

Moreover, the study adopts a grammatical perspective to conduct a qualitative analysis of translation quality, thereby providing additional insights for evaluating machine translation results. It highlights the limitations of machine translation in translating legal text from Chinese to English and offers recommendations for the better design and development of machine translation platforms to meet client needs. Additionally, it presents a perspective for post-editing in machine legal translation from Chinese to English. Overall, this study holds some implications for stakeholders concerning the use and development of machine translation systems for legal texts in the Chinese-English language pair.

## 2. Literature Review

*2.1 Literature Review of Automatic Evaluation of Translation Quality*

The automatic evaluation of translation quality is a crucial task within the fields of natural language processing and machine learning. As such, numerous approaches and methods have been proposed by researchers in recent years to improve the efficiency and accuracy of translation systems. One of the earliest and commonly used methods for automatic translation quality evaluation is BLEU (bilingual evaluation understudy), which relies on n-gram matching between the reference and machine-generated translations (Papineni et al., 2002). While effective, BLEU has been criticized for its inability to capture semantic and syntactic differences between translations. To address this, variations of BLEU have been introduced, such as METEOR (Metric for Evaluation of Translation with Explicit Ordering), which is a syntactic-based evaluation metric that utilizes a combination of matching and stemming (Lavie & Agarwal, 2007). This approach has proven to be particularly advantageous for languages that require a semantic and syntactic perspective. Recently, neural network-based models have been introduced, such as NQE (Neural Quality Estimation), which employs multilingual sentence encoders to learn a function that measures machine translation quality (Niehues & Waibel, 2018). NQE has shown state-of-the-art performance in several evaluation campaigns and has been widely adopted by the machine translation research community.

In conclusion, automatic evaluation of translation quality has become an active area of research, aimed at providing efficient and objective evaluation of translation systems. While traditional metrics such as BLEU score have been widely used, newer metrics that incorporate semantic and syntactic models and neural network-based methods have shown promising results. Future work in this field will likely focus

on developing more sophisticated metrics that can capture the nuances of translation quality, as well as on the development of robust evaluation methods that can cope with different types of translations and language pairs.

*2.2 A Structured Review of BLEU*

The BLEU metric has been widely used in NLP for over 15 years to evaluate NLP systems, especially in machine translation and natural language generation. Several studies have been conducted to evaluate the quality and accuracy of machine translation by using automatic evaluation metrics such as BLEU.

BLEU (Papineni et al., 2002) is a metric that is widely used to evaluate Natural Language Processing (NLP) systems which produce language, especially Machine Translation (MT) and Natural Language Generation (NLG) systems. Because BLEU itself just computes word-based overlap with a gold-standard reference text, its use as an evaluation metric depends on an assumption that it correlates with and predicts the real-world utility of these systems, measured either extrinsically (e.g., by task performance) or by user satisfaction. “Evaluating Machine Translation Systems with BLEU” (Papineni, Roukos, Ward, and Zhu). This paper introduced BLEU as a metric for evaluating the quality of machine translation. The authors tested the metric on a number of translation systems and found that it was a reliable way to assess the accuracy of the translations”. A Comparative Study on Evaluation Metrics for Machine Translation” (Wang, Chung, and Chang). In this paper, BLEU was compared to other metrics such as METEOR, GTM and ROUGE. Their study found that BLEU performed well in terms of correlation with human judgment. D. Palmer (2005) introduced User-Centered method to evaluate MT, which is based on comparing MT output to human referenced translation. The evaluation of MT on Arabic to English and Mandarin to English ranks MT output in comparison to referenced huamn translation. BLEU evaluation method is language independent and can be used to asses any natural language. Yang et al. (2008) studied in Extending BLEU Evaluation Method with Linguistics Weight research to improve the effectiveness of BLEU method, and they succeeded through the use of multiple N-gram weights. Condon et al. (2012) evaluated 2-way Iraqi Arabic-English speech translation systems using automated metrics. They found that automatic translation of Iraqi Arabic correlates with human judgement. Moreover, Adly and Ansary (2009) evaluated Arabic machine Translation by using three automatic measures: BLEU, FI and F mean. Their evaluation is based on Universal Networking Language (UNL) and the Interlingua approach for machine translation.

The previous research has highlighted the importance of automatic evaluation. Various scholars have contributed to this by proposing metrics that measure the closeness of MT output to reference texts, ranging from ranking metrics to grader systems. Specifically, the metric known as Bilingual Evaluation Understudy (BLEU) is widely regarded as the most popular and reliable machine evaluation metric for MT. As noted by EuroMatrix, BLEU has been widely adopted and is considered the best-known machine evaluation metric for MT. Furthermore, Al-Kabi et al. (2013) explain that BLEU aims to determine the quality of any MT system by assessing the closeness of the candidate translation output

to the reference translation produced by a human professional.

Papineni et al. (2001) confirmed that BLEU uses n-gram precision to distinguish between strong and weak translations of MT. Vilar et al. (2006) emphasized the importance of comparing the output of MT to existing translations to identify strengths and limitations of MT systems. They suggest that in order to pinpoint errors in translations, one or more reference texts are needed for comparison to contrast the output of MT systems with correct texts. BLEU has been found to produce reliable results, and can be compared to other metrics for further validation. Previous research has shown BLEU to be a valuable tool for evaluating the quality of machine translation in English-language literature, allowing reliable assessment of the quality of translations produced by MT systems. Therefore, this study also adopts BLEU as a reliable way to assess the quality of MT output.

Despite significant advancements in machine translation technology, evaluating the effectiveness of these systems in accurately translating legal texts remains a relatively underexplored area of research. Complex sentence structures, unconventional word order, and specialized legal terminology make legal texts a unique form of language, warranting the need for rigorous assessment of machine translation quality. Thus, it is crucial to conduct in-depth evaluations to determine the optimal online machine translation system for translating legal texts from Chinese to English.

## 3. Description of Translation Practice and Methodology

### *3.1 The Translation Platforms in the Study*

Choosing the suitable machine translation engine is crucial for legal translation. The quality of the translation output produced by the machine engine greatly affects the translator's workload. A high-quality machine translation result would only require minor revisions or corrections from the translator. Thanks to advancements in technology such as neural machine translation and artificial intelligence, machine translation has significantly improved in quality, and many scholars have recognized its benefits. Popular machine translation engines in China include Youdao, Baidu, and Sougou, etc. while foreign ones include DeepL and Google, etc. and all have a long history.

Two popular online translation systems in the legal industry, DeepL (Note 2) and Metasota (Note 3) are selected for this evaluation.

#### 3.1.1 Introduction of DeepL

DeepL is an artificial intelligence translation tool from Germany that utilizes neural networks to provide full-text translation for nine languages, including Chinese, English, and German. Its superior performance has surpassed tech giants such as Google, Microsoft, and Facebook in the machine translation field. In blind tests, DeepL's translation results are favored by a factor of 3:1 over its competitors, and it has achieved record-breaking performance according to scientific benchmarks. Its speed is comparable to its massive competitors, yet its accuracy and detail stand out. Thus, DeepL is at the forefront of machine translation engine development and has been selected as one of the MT engines for legal translation evaluation in this thesis.

3.1.2 Introduction of Metasota

Metasota is an emerging technology company that specializes in legal artificial intelligence. Their goal is to improve efficiency in the legal industry by creating artificial intelligence tools for legal professionals. Their services have been widely adopted by major law firms and corporate legal departments, with strong performance in translating contracts, legal documents, and policy correspondence. To evaluate machine translation performance in legal texts, we compare the popular foreign online translation software DeepL with domestic Metasota for translating Chinese to English.

Both systems are freely accessible and able to handle long sentences of at least 100 words, which is necessary in legal texts. This standard eliminates online machine translation systems that impose strict word count limits. (Note 4)

*3.2 The Corpus Used in the Study*

The quality of machine translation output is heavily dependent on the quality of the source text. Grammatically correct and concise input facilitates accurate translations, while misspellings, slang, and convoluted syntax can impede the process. Legal texts, with their lengthy and complex sentences, specialized terminology, and irregular word order, require linguistically appropriate input. To evaluate two machine translation systems fairly in the context of legal translation, a reliable test set is needed. For this reason, Law of the People's Republic of China on Combating Telecom and Online Fraud (Note 5) which was adopted by the 36th session of the Standing Committee of the 13th National People's Congress on September 2, 2022, is chosen as the test text, and the English translation of this law, which was completed by the professional translation team (pkulaw) on September 5, 2022, is chosen as the reference translation for the BLEU evaluation. The reference translation for comparison. Each language has a high standard and is translated by professionals. Therefore, this test text is used to examine the quality of machine translation systems for the translation of legal texts and to generate statistical analyses for their performance in quantitative evaluations, a reliable test set.

In this study, since the free version of Metasota can only translate texts less than or equal to 100 words at a time, 25 of these articles were selected (in Chinese as the source language, totaling as 1495 Chinese characters) in order to compare the translation performance of DeepL and Metasota more fairly. And they contain simple sentences and complex long and difficult sentences, so that the test results are more reasonable and fair.

*3.3 Choosing BLEU Tool*

Within the domain of Machine Translation (MT), a vital requirement is to develop a reliable indicator that enables an accurate measurement of MT output quality. In traditional evaluation methods, human assessment is a commonly utilized approach for evaluating translation output quality, which considers numerous aspects, such as adequacy, fidelity, and fluency. However, relying solely on human evaluation can be time-consuming and labor-intensive. Consequently, there is a pressing need for an efficient method to evaluate MT output.

To address these issues, researchers in the MT field have invented automatic evaluation metrics, such

as BLEU, METEOR, and NIST, to generate more objective and reliable evaluation results. The BLEU method developed by Kishore Papineni is one such metric, which essentially measures the degree of similarity between the computed MT output and human translation output. The fundamental calculation methodology relies on comparing the number of common words and word sequences, known as n-grams, found between candidate and reference translations.

Applying BLEU entails inputting the MT output into the metric, which then calculates the n-gram metric to generate a score between 0 to 1. The closer the score is to 1, the higher the quality of the MT output. BLEU has gained prevalence in evaluation metrics because it displays the highest correlation with human judgment. As a result, it is frequently employed in MT research and evaluations, that is why the BLEU measure is used in this study. To summarize, the BLEU metric is becoming increasingly important in the field of MT as it enables a more objective and reliable assessment of translation quality.

## 4. Research Procedure

The research process of this study is mainly divided into two parts, the first part is the automatic evaluation process used in this study, we choose the BLEU program as the automatic evaluation method, and compare the output quality of two machine translation platforms in translating legal texts from Chinese to English by analyzing the BLEU scores of the two machine translation platforms, the second part is the manual evaluation and analysis process all in this study, from the perspective of grammar including the analysis of The second part is the manual evaluation and analysis process of this study, which includes the analysis of terminology usage accuracy, word order, subject-predicate collocation, sentence structure, tense and morphology, etc. from the grammar perspective. to manually evaluate the machine translation output, and further summarize and analyze the output errors and shortcomings of the two machine translation platforms. The following two sections will introduce this part in detail.

### *4.1 The Automatic Evaluation*

The authors first visited the websites of the candidate machine translation systems and collected information on the language pairs they cover, and then we analyzed the evaluation texts of the candidate machine translation systems.

Twenty five articles were obtained by extracting texts from web pages in these languages, and the resulting Chinese text was used as the source text, whose English reference translation was used as the reference standard for evaluation, which was performed by comparing the translation results of each candidate machine translation system for the source text with the reference translation. The source text was submitted to one of the two selected machine translation systems sentence by sentence, and then the English translation results of its output were collected.

Then, the author converted the PDF file into TXT format for BLEU to discern in the flowing step. Then, the author screened and removed the special symbols in the source text, because BLEU could not

recognize the special symbols. Once the special symbols appeared, BLEU would report errors. After that, the author aligns source text with reference translation. Secondly, the study put the source text into DeepL Translate and Metasota Translate respectively. Through machine translation, the author got two machine translation outputs.

Thirdly, the study aligned two machine translation output files with the reference translation respectively, because BLEU will report the error in cases of misalignment in the next step.

At last, the author put two machine translation files and reference translation in to the Python BLEU program respectively for calculation to verify a valid and optimal system for legal text translation and obtain the BLEU score. BLEU is a 0-1 scale in which higher score means greater similarity between the machine translation and reference translation.

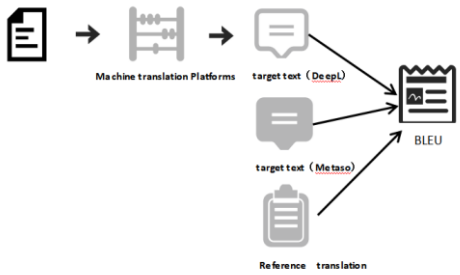



**Figure 1. Flow Chart of the Preliminary Work of This Study**

The author analyzed these sentences to find N-gram strings, and in this regard, BLEU used the N-gram (Note 6) matching rule, by which it was able to calculate a percentage of similarity of n groups of words between the comparison translation and the reference translation thus in order to calculate the accuracy of each output translation compared to the human translation of the reference translation.

The present section presents the analysis and measurement of the MT output to conduct this study to verify the effectiveness of MT output through text similarity metric, BLEU.

To explain how BLEU works, the following example elucidates how to measure the closeness between the candidate (MT output) and human referenced translation. Our example is:

Taking Article 6 of the Anti- Fraud Act as an example, we see the matching of 1-3 gram as follows:

Source text：

国务院建立反电信网络诈骗工作机制，统筹协调打击治理工作。

The selected source text was fed into the DeepL and Metasota online machine translation systems and the output was as follows:

**Table 1. The Output of the Selected Source Text**

| |
|---|
| DeepL:<br>The State Council established an anti-telecommunications network fraud working mechanism to coordinate efforts to combat and manage the problem. |
| Metasota:<br>The State Council shall establish a working mechanism for anti-fraud committed through telecommunication networks and coordinate the crackdown and control work as a whole.<br>Reference:<br>Article 6 The State Council shall establish a working mechanism for combating telecom and online fraud and conduct the overall coordination of crackdown and governance. |

Unigram

We can see that there are 23 words in the reference translation and 19 words in the DeepL machine translation, of which 7 words are the same as the reference translation, so its match is 7/19.

Bigram

Similarly, the match of the binary phrase is 3/18.

Trigram

The ternary phrase has a match of 1/17.

In general, Unigram can represent how many words of the original text have been translated individually, which can reflect the adequacy of the translation, and Bigram or above can reflect the fluency of the translation, and the higher the value of it indicates the better readability. These two indicators are able to be benchmarked against the manual evaluation. The closer the words and ordering are to the reference translation, the higher the accuracy score is, and the N-gram results about this example sentence are shown in Tables 2 and 3.

**Table 2. Accuracy Value**

| MT system / N-gram | DeepL | Metasota |
|---|---|---|
| Unigram | 7/19 | 13/24 |
| Bigram | 3/18 | 9/23 |
| Trigram | 1/17 | 7/22 |

**Table 3. N-Gram Precision Value**

| MT system / N-gram | DeepL | Metasota |
|---|---|---|
| Precision (1) | 0.37 | 0.54 |
| Precision (2) | 0.17 | 0.39 |
| Precision (3) | 0.06 | 0.32 |

Using the above calculation steps, the author calculated the output performance of Chinese to English translations of the two machine translation systems in twenty five test texts in turn. The average values are as follows.

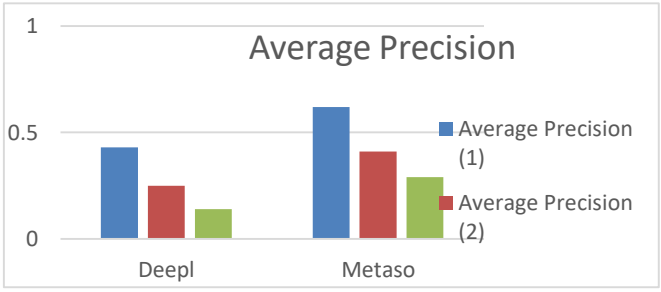



**Table 4. Average N-Gram Precision**

However, the N-gram accuracy value also has its limitations. If the output of machine translation is short sentences, the accuracy of computing n-gram will score high, but in fact it should score lower. For the case that the length of the translated translation is shorter than the reference translation, a penalty mechanism (Brevity Penalty) is needed to control it.

Penalty factor, Brevity Penalty BP

Secondly, we calculated the score of BLEU by computing Brevity Penalty value by choosing the best system.

The author analyzed the sentences to fine the N-gram strings to calculate the precision of each sentence in comparison to human referenced translation. In this respect, the IBM formula of BLEU is adopted to measure the precision as follow:

$$\mathrm{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

c=word length of the machine translation output

r=word length of the reference translation

The penalty rule is that if the candidate translation is longer than the reference translation, then BP is set to (1), while if it is shorter, the result is (0).

In our example, for DeepL, c = 19 and r = 23. When 19 < 23, BP = 1; for Metasota, c = 24 r = 23 When 24 < 23, BP also = 1. Then, we substitute the Brevity Penalty (BP) result in the first equation to calculate the final BLEU score, and the BP value penalizes those candidate sentences that are shorter than the reference translation The accuracy of the BLEU measure ranges from 0 to 1, where the closer the score is to 1, the closer the translation is to the reference text.

After various improvements above, the final formula of BLEU is as follows.

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right).$$



DeepL：BLEU=BP*(1/3*log(7/19)+1/3*log(3/18)+1/3*log(1/17))=0.34

Metasota：BLEU=BP*(1/3*log(13/24)+1/3*log(9/23)+1/3*log(7/22))=0.43

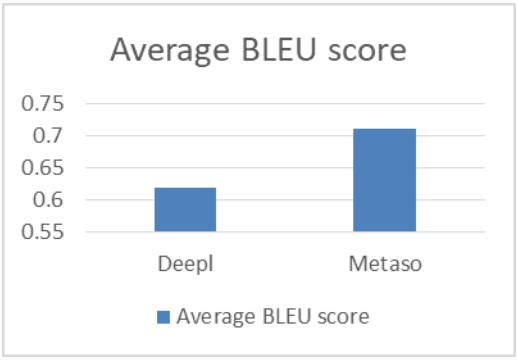
The author then used this method to calculate 25 test texts, and obtained the data as follows (the results retain three decimal places).

**Table 5. The BLEU Score of the 25 test Texts**

| Source text | DeepL Translation | Metasota Translation |
|---|---|---|
| 1 | 0.708 | 0.666 |
| 2 | 0.707 | 0.767 |
| 3 | 0.619 | 0.687 |
| 4 | 0.664 | 0.752 |
| 5 | 0.618 | 0.727 |
| 6 | 0.573 | 0.671 |
| 7 | 0.615 | 0.656 |
| 8 | 0.765 | 0.581 |
| 9 | 0.378 | 0.596 |
| 10 | 0.646 | 0.678 |
| 11 | 0.711 | 0.634 |
| 12 | 0.752 | 0.876 |
| 13 | 0.615 | 0.86 |
| 14 | 0.705 | 0.815 |
| 15 | 0.565 | 0.704 |
| 16 | 0.607 | 0.653 |
| 17 | 0.528 | 0.723 |
| 18 | 0.640 | 0.720 |
| 19 | 0.479 | 0.581 |
| 20 | 0.541 | 0.679 |
| 21 | 0.429 | 0.592 |
| 22 | 0.573 | 0.671 |
| 23 | 0.609 | 0.791 |
| 24 | 0.743 | 0.767 |
| 25 | 0.948 | 0.884 |

From this, the author calculated the average values of BLEU for the two machine translation systems in twenty five test sentences, and the results are shown in Table 6.

**Table 6. Average BLEU Score of the 25 test texts**




The above graphs show that the translation performance of Metasota and DeepL for legal texts with 1-3gram values and BLEU averages are higher than those of DeepL. The above graphs indicates that the text similarity between the Metasota and the human reference translation is higher and therefore the translation output is more reliable.

Results:

The full results of the survey are presented in the data file associated with this article. I summarize key findings here.

One important question is what level of correlation is sufficient for BLEU to be regarded as a valid proxy for human evaluation. I will use the following classification.

High: Correlation is 0.85 or higher

Medium: Correlation is between 0.70 and 0.85

Low: Correlation is between 0 and 0.70

The High, Medium, and Low classification is based on the classification of surrogate endpoints in Prasad et al.(2015), which in turn is based on criteria from the Institute for Quality and Efficiency in Health Care (IQWiG) in Germany, which assesses the benefits of medical interventions.

## 5. Research Findings

Translation from Chinese to English is a challenging task due to the syntactic, morphological, and semantic differences between the two languages. This study evaluates the bilingual test text of the "Law of the People's Republic of China on Combating Telecom and Online Fraud" extracted from pkulaw.

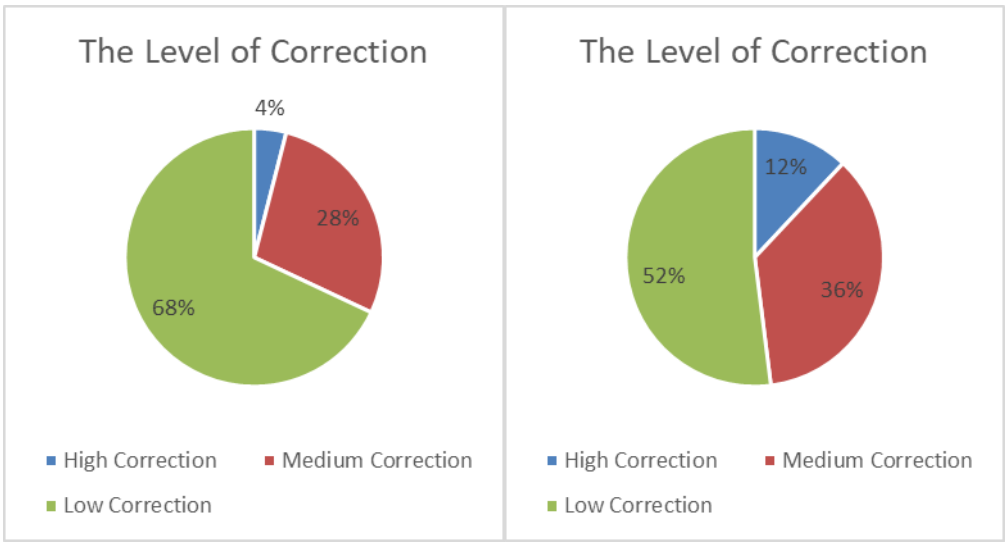
The study employs the BLEU metric, a real and authoritative qualitative evaluation method in the field of machine translation in recent years, to empirically evaluate the accuracy values of the Chinese-to-English output translations of two online machine translation systems compared with the human reference translations. This practical machine translation evaluation method incorporates human judgment and uses human translation as a reference, and therefore provides more reliable, objective and consistent evaluation results compared to human evaluation. Selecting a suitable machine translation

system for users when translating legal texts from Chinese to English.

The results of the study show that the average BLEU score of DeepL's 25 test texts is 0.62 (two decimal places are retained), and one text has high relevance in BLEU value (4%), seven texts have medium relevance in BLEU value (28%), and 17 texts have low relevance in BLEU value (68%). The average BLEU score of Metasota translation is 0.71 (two decimal places are retained). And there are 3 articles with high correlation of BLEU value, accounting for 12%; 9 articles with medium correlation of BLEU value, accounting for 36%; 13 articles with low correlation of BLEU value, accounting for 52%. And as shown in Table 7.

This result indicates that Metasota achieved better results than DeepL translation system, and Metasota translation has the highest text similarity with human reference translation and higher correlation with human reference translation. This is because a higher BLEU score for any machine translation means that its translation output is more reliable than other translation systems. The results show that Metasota is more effective and reliable than DeepL in translating Chinese legal texts into English. This information is needed by potential users. However, because the high correlation values of these two online machine translation BLEU values are still small, it shows that online machine translation systems are still far from achieving the overall quality of human translation. In the foreseeable future, although machine translation systems can assist human translation of legal texts to a large extent, it is impossible for machine translation to completely replace human in the translation of legal texts, and human post-translation editing is still required. This study proposes to analyze machine translation output translations from a grammatical perspective to provide some information in order to help system designers focus on where they are temporarily lagging behind and promote the development of machine translation.

**Table 7. The Level of Correction of Deepl's outputs and Metasota's Outputs**




DeepL Metasota.

## 6. Error Analysis

This chapter intends to give a detailed analysis of the existing errors of the two online automatic MT systems of DeepL and Metasota, especially the DeepL. The existing errors are focused on at 6 levels, briefly speaking, the accuracy of terminology usage, word order, subject-verb agreement, sentence structure, tense, and voice.

### *6.1 Accuracy of Terminology*

1) For example, in the translation of Article 3, Metasota uses more accurate legal translation terms to express, for example: “within the territory of” “by citizens of” such expressions, which more accurately indicate the location and subject of online fraud, making the translation more accurate and more in line with the expression habits of legal terminology, and in full compliance with the expressions in the reference translation, making the translation of Metasota more relevant to the human translation.

2) In the translation of Article 5, DeepL directly translates the word “进行” as “carried out”, while both the reference translation and the Metasota translation are more accurately translated as “conducted”.” Carry out” usually refers to the implementation of a plan or task, and the word “conduct” here enriches the meaning of the phrase. The same is true of DeepL, which translates the word “维护” as “protect”, while both the reference translation and Metasota translate it as “safeguard”. The meaning of “维护” is better reflected.

3) In the translation of Article 5, both the Metasota translation and the reference translation add “fraud” between “fight” and “commit” to more accurately express the meaning of “诈骗”. It is more accurate than the translation of DeepL, which expresses that this is a kind of fraud. It is more accurate than the translation of “DeepL”, which meets the requirements of legal terms.

4) In the translation of Article 5, Metasota’s translation’s tone is stronger than DeepL’s translation because it uses the phrase “shall be conducted”, which expresses responsibility and obligation, and in legal texts, shall is often used to express obligation, responsibility and instruction. Therefore, Metasota’s use of shall is more in line with the expression habits of legal translation, expressing the strong will to combat online fraud and protect the legal rights and interests of citizens and organizations in accordance with the law, while DeepL uses the word should, which has a slightly euphemistic tone, expressing mainly advice, counsel, recommendation and other meanings, which is not as strong as Metasota’s expression and does not conform to the expression habits of legal documents.

5) In the translation of Article 20, Metasota uses more appropriate legal terminology, and since the translation is for legal documents, the accuracy of the terminology is crucial. Metasota’s translation of “quick freezing” is more in line with legal terminology, while DeepL uses “rapid freezing”, which is not accurate.

### *6.2 Word Order*

When it comes to legal translation, one of the very important factors is the order of words. Since the order of words may vary greatly from language to language, care must be taken to maintain the similarity of sentence structure and order in legal texts when translating to ensure that the information

conveyed remains accurate and complete.

1) In the translation of Article 3, Metasota's word order more closely resembles English expressions. For example, in English, the subject is often placed at the beginning of the sentence, followed by the verb and the object. In Metasota's translation, the subject " fraud" is placed at the beginning of the sentence, followed by the verb phrase "committed through telecommunication networks" and then the prepositional phrase "within the territory of the People's Republic of China" or "by citizens of the People's Republic of China Outside the territory of the People's Republic of China" finally indicates the scope of application of the law, while the translation of DeepL is not quite in line with the English language pattern, using the parallel conjunction, and, to connect two different information This makes the sentence structure complicated and more difficult to understand. Therefore, in legal translations, the use of expressions that conform to the order of the original language can make the translation more readable and convey the legal content more accurately.

2) In the translation of Article 35, the translation of Metasota is more reasonable in terms of language order and logical arrangement, and can better present the information of the original language. Specifically, Metasota's translation uses the phrase "upon the decision or approval of the anti-fraud committed through telecommunication networks under the State Council" accurately expresses the meaning of "经国务院反电信网络诈骗工作机制决定或者批准", and places it at the beginning of the sentence, which emphasizes the importance of the decision or approval, better presents the information and intention of the original text, and conforms to the expression custom of legal translation.

*6.3 Subject-Verb Agreement*

Metasota's translation is generally more reasonable than DeepL's translation in terms of subject and predicate, and can be consistent with the subject-predicate collocation of the reference translation.

1) For example, in the translation of Article 20, Metasota uses "the public security department of the State Council shall" as the subject, and "establish and improve" as the predicate. This expression is clear, and follows the expression habit of the English official documents. The translation of DeepL uses "public security departments of the State Council" as the subject, and the predicate after it uses the singular form of "to establish "This is not in accordance with the English grammar.

2) In the translation of Article 48, the subject and predicate of Metasota are more accurately matched, and DeepL's use of the expression "you can apply" seems to regard the behavior of the subject (relevant units and individuals) as a suggestion or choice. But in fact, the legal rights and obligations of citizens, legal persons or other organizations are stipulated in the legal text, and administrative reconsideration and administrative litigation are one of the remedies. Therefore, Metasota uses the term "may apply" to more accurately express the legal provisions.

*6.4 Sentence Structure*

1) For example, in the translation of Article 2, the sentence structure of Metasota is very similar to the reference translation, and the sentence structure is clearer, and the information is expressed in three clauses according to "打击治理", "境内实施" and "境外实施". "The information is expressed in three

separate clauses, which makes the information in each part more accurate and avoids the difficulty of comprehension due to long sentences. The translation of DeepL, on the other hand, uses "and" to connect the two parts of information, resulting in a complex grammatical structure, which makes it more difficult for readers to understand.

2) In the translation of Article 5, the relationship between the subject "the fight against fraud committed through telecommunication networks" and the predicate "shall be conducted" in the translation of Metasota are more closely related. In contrast, DeepL's translation has a longer definite clause (shall be carried out in accordance with the law) between the subject (anti-telecommunications network fraud work) and the predicate (shall be carried out) (should be carried out in accordance with the law), resulting in the subject-predicate relationship is not compact.

3) In the translation of Article 22, from the grammatical point of view, the translation of Metasota is better, because the translation of Metasota reflects the subject-verb-object structure in the sentence, which makes the sentence clearer, while the translation of DeepL adopts the structure of "gerund + subject + predicate", which makes the relationship between the subject and the predicate more The relationship between the subject and the predicate is blurred. In legal documents, complex language structures should be avoided to avoid ambiguity and misunderstanding.

4) In the translation of Article 44, Metasota uses more accurate and standardized grammatical structures, such as "if...then" conditional sentences, which is more in line with the expression habits of legal texts.

5) In the translation of Article 48, Metasota's nouns and verbs are more appropriate, and DeepL's use of "administrative compulsory measures decisions" may make the reader think that the decisions themselves are "强制措施" rather than decisions on "强制措施". However, in the original text, "administrative compulsory measures" itself is a noun phrase, and "decisions" is its definite article, therefore, Metasota's The translation of Metasota better preserves the structure of the original.

*6.5 Tense and Voice*

1) For example, in the translation of Article 22, the translation of Metasota uses the present perfect tense and passive voice, which makes the sentence more accurate and standardized, compared with the translation of DeepL, which uses the past tense, may cause confusion and misunderstanding among readers.

2) In the translation of Article 44, Metasota's translation uses the passive voice of "shall be imposed", which is more in line with the expression habits of legal texts.

In addition to grammar and expression habits, legal translations also need to follow some principles, such as accuracy, clarity and completeness. Metasota's translation is generally more standardized and better reflects the principle of standardization and rigor in legal translation.

For example:

3) In Article 2, Metasota uses quotation marks to highlight the term "fraud committed through telecommunication networks", which is in line with the standard requirement of legal translation for

practical terminology, while DeepL's translation does not use quotation marks and fails to highlight the term, which is not standardized.

4) In Article 23, Metasota's translation uses the word "any" to express the universality of the licensing or filing procedures, which can more accurately convey the meaning of the original text. In addition, Metasota uses the expression "telecommunication department", which is more specific than "telecommunication authorities". clearer and easier to be understood by readers.

## 7. Conclusion

### *7.1 Conclusion and Future Prospects*

To sum up, in this study, we selected the newer legal texts in China as the real data objects, and we concluded that the translation output of Metasota is more reliable in correlation with the human reference translation through both measured BLEU score performance of DeepL and Metasota Chinese to English legal text output. After that, through qualitative analysis, from the perspective of grammar (specifically, the accuracy of terminology, word order, subject-predicate collocation, sentence structure, tense and morphology), taking into account the principles of legal translation and the characteristics of legal translation, we analyzed what Metasota's translation output did better than DeepL's translation output: compared with DeepL's translation, Metasota's translation selected more accurate legal terms, had more reasonable word order, and had more reasonable sentence structure. Compared with the DeepL translation, the Metasota translation is more accurate in the choice of legal terms, more reasonable in the arrangement of language order, more standardized in the structure of sentences, more reasonable in the collocation of subjects and predicates, and closer to the quality of human translation. It is a better choice for legal workers and legal translation enthusiasts. Using Metasota online machine translation system can greatly reduce the time of post-translation editing and improve the accuracy of the translation at the same time. At the same time, we can also recognize that in the field of legal text translation, there is still a considerable gap between machine translation and human translation, and the current machine translation system is still difficult to achieve high-quality legal text translation, unable to meet the real demand, let alone replace human translation. In order to improve machine translation for legal translations, there are several areas that need to be addressed:

1) Accuracy: Machine translation systems need to improve their accuracy, especially when dealing with legal terminology which often contains specific meanings and nuances.

2) Contextual understanding: To accurately translate legal texts, machine translation systems must have the ability to understand the context in which specific words are used in legal documents.

3) Customization: Customizing machine translation systems for specific legal domains, such as contract law or intellectual property law, will enable more accurate and specialized translations.

4) Natural language processing: Future machine translation systems must improve their natural language processing capabilities in order to produce translations that are more natural and easier to comprehend.

5) Ethical considerations: Machine translation providers need to consider the ethical implications of their technology, particularly when it comes to confidentiality and privacy concerns in legal translations.

Therefore, in the future, there is still a need to strengthen the continued cooperation between machine translation and human post-translation editing, thus improving the accuracy of legal text translation.

*7.2 Research Limitations*

First of all, BLEU assessment has its own limitations, such as: it does not consider the accuracy in linguistic expressions (grammar); the accuracy of the measurement can be disturbed by commonly used words; the accuracy of the measurement is sometimes higher for short sentences; and it does not consider synonyms or similar expressions, which may lead to the negation of reasonable translations. Also the BLEU evaluation process lacks linguistic knowledge, e.g., a large-scale human post-editing analysis from a Google research team pointed out that machine translation outputs that are qualitatively worse than human-edited translations occasionally have BLEU values of up to 5 points.

Second, only a single text type, twenty five test sentences of a legal text, was used in this evaluation, and translations were limited to Chinese-to-English language pairs; a more comprehensive evaluation of more machine translation systems and more languages with more texts would undoubtedly provide a more complete picture of the full potential of online machine translation systems.

In terms of the subsequent theoretical analysis, no linguistic theory that fits well with the comparison of Chinese to English translations of legal texts by two machine translation systems was found, and the final choice was to start the analysis in terms of grammar and the principles of legal translation. Therefore, the lack of theoretical support is also a shortcoming of this paper.

**Reference**

Al-Kabi et al. (2013). Evaluating English to Arabic machine translation using BLEU. *International Journal of Advanced Computer Science and Applications*, *4*(1). https://doi.org/10.14569/IJACSA.2013.040109

Almahasees, Z. M. (2017). Assessing the translation of Google and Microsoft Bing in translating political texts from Arabic into English. *International Journal of Languages, Literature and Linguistics*, *3*(1), 1-4. Information Systems, 2009, pp. 243-257. https://doi.org/10.18178/IJLLL.2017.3.1.100

Bahdanau, D., Cho, K., & Bengio, Y. (2014). *Neural machine translation by jointly learning to align and translate*. arXiv preprint arXiv:1409.0473.

Chen, J., & Dai, D. (2019). Evaluation of Machine Translation in Cross-border E-commerce based on BLEU. *Journal of the China Society for Scientific and Technical Information*, *38*(11), 1286-1297.

Han, L. (2022). *An Overview on Machine Translation Evaluation*. arXiv preprint arXiv:2202.11027.

Hutchins, W. J., & Somers, H. L. (1992). An introduction to machine translation. *Academic press*.

Jebbar, A. Machine Translation Vs Human Translation. (n.d.). Retrieved from

http://www.translationdirectory.com/articles/article1326.php

K. Papineni et al. (2001). BLEU: A method for automatic evaluation of machine translation. In *Proc. the 40th Annual Meeting on Association for Computational Linguistics, Stroudsburg, PA, USA*, 311-318. https://doi.org/10.3115/1073083.1073135

Kit, C., & Wong, T. M. (2008). *Comparative evaluation of online machine translation systems with legal texts*. Law Libr. J., 100, 299.

Koehn, P. (2010). *Statistical machine translation*. Cambridge University Press. https://doi.org/10.1017/CBO9780511815829

M. Yang et al. (2008). Extending BLEU evaluation method with linguistic weight. In Proc. *the 9th International Conference for Young Computer Scientists* (pp. 1683-1688). https://doi.org/10.1109/ICYCS.2008.362

Ma, J., Wu, L., Jiang, Y., & Wang, J. (2019). Quality Evaluation of Machine Translation for Legal Language based on BLEU. *Computer-Assisted Foreign Language Education*, *2*(01), 42-49.

N. Adly., & S. Ansary (n.d.). *Evaluation of Arabic machine translation system based on the universal networking language*. in Proc. the 14th International Conference on Applications of Natural Language.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). 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). https://doi.org/10.3115/1073083.1073135

Reiter, E. (2018). A structured review of the validity of BLEU. *Computational Linguistics*, *44*(3), 393-401. https://doi.org/10.1162/coli_a_00322

Schiaffino, R., & F. Zearo. (n.d.). *Translation Quality Measurement In Practice*. Retrieved from http://www.Translationquality.com

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). *A study of translation edit rate with targeted human annotation*. Proceedings of association for machine translation in the Americas, 223-231.

Wang, X., & Huang, Y. (2020). Quality Evaluation of Legal Translation by Machine Translation: Current Status and Challenges. *Chinese Journal of Law*, (6), 141-151.

Wanghao, G., & Fumao, H. (2021). A Study of the Assessment of Translations and Post-editing in Neural Machine Translation. *Journal of Beijing International Studies University*, *43*(5), 66.

Wei, Y. W., Li, N., & Zhao, L. W. (2022). Machine translation quality standards based on high frequency error type analysis. *Computer Science and Application*, *12*, 2275. https://doi.org/10.12677/CSA.2022.1210232

Yu, X. H. (2020). *Comparative quality study of professional machine translation and General machine translation*.

Zhang, X. J. (2007). Review of the quantitative evaluation of translation quality. *Foreign language research*, (4), 80-84.

Zhang, Q., & Liu, Y. (2020). Evaluation Research on Machine Translation Quality of Legal Texts based on BLEU Method. *Computer Simulation*, *37*(5), 246-251.

Zhao, Y., & Chen, X. (2020). Research on Evaluation Model of Legal Translation Quality based on Improved BLEU. *Computer Engineering and Applications*, *56*(14), 93-100.

Zhou, G. F., & Gao, Y. J. (n.d.). *On the assessment of machine translation quality*.

**Notes**

Note 1. Automatic Language Processing Advisory Committee, Language and Machines: Computers in Translation and Linguistics (1966) (Publication 1416, National Academy of Sciences National Research Council), available at http://www.nap.edu/openbook.php?isbn=ARC000005.

Note 2. DeepL, https://www.DeepL.com/translator (last visited Apr. 10, 2023).

Note 3. Metasota, https://fanyi.Metasota.cn/#/ (last visited Apr. 10, 2023).

Note 4. For example, Reverso (www.reverso.net/text-translation.asp) has a limit of 400 characters per translation, while Amikai (www.amikai.com/demo.jsp) has a limit of 100 characters.

Note 5. Law of the People's Republic of China on Combating Telecom and Online Fraud is a regulation enacted in accordance with the Constitution in order to prevent, curb and punish telecommunication network fraud activities, strengthen anti-telecommunication network fraud, protect the legitimate rights and interests of citizens and organizations, and maintain social stability and national security. Law of the People's Republic of China on Combating Telecom and Online Fraud has seven chapters and 50 articles, including general provisions, telecommunications governance, financial governance, Internet governance, comprehensive measures, legal liability, and bylaws, adhering to the people as the center, coordinating development and security, based on all aspects and the whole chain of prevention and management of telecommunications network fraud, and precise efforts to provide strong legal support for the work of anti-telecommunication network fraud. 2 September 2022. The Thirty-sixth Session of the Standing Committee of the Thirteenth National People's Congress voted to adopt the "Law of the People's Republic of China on Combating Telecom and Online Fraud", which came into force on December 1, 2022.

Note 6. N-gram is a statistical language model, N-gram means "N-gram", which refers to N consecutive words in a text, and is a tool in statistical language models. When N equals to 1, we call it 1-gram or Unigram, when N equals to 2, we call it 2-gram or Bigram, and when N equals to 3, we call it 3-gram Trigram, which can represent a sentence as a sequence of n consecutive words, and use the collocation information between adjacent words in the context to calculate the probability of the sentence, so as to determine whether a sentence is fluent or not.