ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 14, No. 9, pp. 2823-2833, September 2024
DOI: https://doi.org/10.17507/tpls.1409.18

A Comparative Study on the Quality of English-Chinese Translation of Legal Texts Between ChatGPT and Neural Machine Translation Systems*

Lijie Ding
School of Foreign Languages, Southwest University of Political Science and Law, Chongqing, China

***Abstract*—This study conducts a comparative analysis of the quality of English-to-Chinese (E-C) and Chinese-to-English (C-E) translation of legal texts between Chat Generative Pre-trained Transformer (ChatGPT) and four online Neural Machine Translation (NMT) systems. The analysis includes both quantitative and qualitative evaluations. The results suggest that both ChatGPT and the NMT systems achieve satisfactory performance in translating legal texts from Chinese to English. Although the quality of ChatGPT's C-E legal translation is slightly lower than that of the NMT systems, the difference is not statistically significant. However, neither ChatGPT nor the NMT systems meet a passing standard for E-C translation of legal texts, with the NMT systems showing better overall performance. Overall, ChatGPT and the NMT systems perform better at translating legal texts from Chinese to English compared to E-C translation. For E-C legal translation, ChatGPT's quality is lower compared to the NMT systems. While the types of errors are similar in both systems, ChatGPT tends to exhibit more errors, some of which are more severe. This study serves as a reference for those choosing translation tools for E-C and C-E legal texts.**

***Index Terms*—ChatGPT, legal translation, Neural Machine Translation**

I. Introduction

Chat Generative Pre-trained Transformer (ChatGPT) is a sophisticated AI chatbot created by OpenAI, an American research laboratory specializing in artificial intelligence. Launched in November 2022, ChatGPT integrates a variety of natural language processing features, including answering questions, telling stories, writing emails, debugging code, and translating texts. This raises the question: How effective is ChatGPT in translation tasks? Specifically, when compared to Neural Machine Translation (NMT) systems, which are also based on AI technology, does ChatGPT offer better performance in specialized translation areas, such as legal translation?

Several scholars have investigated the translation quality of ChatGPT from different perspectives. Khoshafah (2023) compared ChatGPT's accuracy in translating texts from various fields such as media, literature, science, religion, and law. She found that while ChatGPT typically provides precise renditions, it exhibits limitations for certain types of texts, including legal papers, medical records, scientific researches, and literary creations. In another study, Jiao et al. (2023) conducted a comparison between ChatGPT and commercial machine translation services. They discovered that ChatGPT performed on par with commercial offerings such as Google Translate in resource-rich European languages but lagged significantly behind in resource-poor languages. Sanz-Valdivieso and Lopez-Arroyo (2023) demonstrated that ChatGPT produced fewer terminology errors compared to Google Translate. Cady et al. (2023) aligned sentences derived from samples sourced from a vast Chinese-English bilingual patent collection and various references. Their results indicated that although ChatGPT exhibited superior performance overall, it did not surpass NMT in every aspect. Hendy et al. (2023) reported that ChatGPT offers high translation quality for high-resource languages but only mediocre quality for low-resource languages. Additionally, Grimm et al. (2024) conducted research on GPT-4 and discovered its ability to generate precise, comprehensible, and practical results in English, Spanish, and Mandarin. Lee (2023) highlighted that although Large Language Models, such as ChatGPT, were not primarily developed for translation tasks, they have demonstrated a level of technical advancement capable of producing renditions that compete with or surpass specialized translation platforms available in the market, such as Google Translate and DeepL.

Several scholars have undertaken comparisons between the translation efficacy of ChatGPT and NMT. Yang (2023) employed ChatGPT to translate Vietnamese legal texts, concluding that ChatGPT exhibited no discernible superiority when juxtaposed against other machine translations and human counterparts. Likewise, Zhao et al. (2023) acknowledged the commendable strides ChatGPT has made in natural language processing, problem comprehension, and user interaction. Yu (2024) found that the syntactic complexity of ChatGPT translation is on par with both human

* This article was supported by the 2023 Student Scientific Research Innovation Project of Southwest University of Political Science and Law under Grant 2023XZXS-321.

translation and DeepL translation. However, it excels beyond human translation in terms of the frequency of coordinate phrases and the intricacy of verb phrases, while also surpassing DeepL translation in average T-unit length.

In recent years, numerous scholars have delved into research on NMT, striving to enhance machine translation (MT) quality through technological advancements. The research findings of Feng and Zhang (2022) highlight that NMT has transitioned into a large-scale practical phase. Particularly in English-Chinese translation, the accuracy rate for general texts has surpassed 90%, effectively meeting the demands for translating everyday materials such as news reports, product instructions, and traffic information. Li (2021) observed that while the translation quality of the five online machine translation (OMT) systems based on neural network technology reached an acceptable threshold, it fell short of achieving a superior or excellent level. These studies collectively underscore the significant strides made in the quality of translating commonplace texts from English to Chinese through NMT.

The above research demonstrates that the performance of ChatGPT and NMT in translation is indeed impressive. However, the corpora used by the researchers were general corpora, and the translation directions involved mutual translation into multiple languages. So far, no one has focused on ChatGPT's performance in English-Chinese legal translation, nor has there been a comparison between ChatGPT and NMT regarding the quality of legal translation.

As globalization progresses, the demand for legal translation between English and Chinese is on the rise. ChatGPT and NMT are among the most advanced translation technologies available today, and conducting a comparative analysis of their respective strengths and weaknesses can offer insights and recommendations for improving translation systems. Moreover, by comparing the quality of their translations, one can gain a clearer understanding of these technologies' capabilities, providing legal translators with useful information for selecting and utilizing translation tools.

This study will compare the performance of ChatGPT-4, with four NMT systems—Youdao Translate, Baidu Translate, Google Translate, and DeepL Translate—to assess their effectiveness in translating legal texts between English and Chinese. The research aims to answer the following questions:

(1) Which system—ChatGPT or NMT—delivers better performance in translating legal texts between English and Chinese?

(2) Using the same evaluation metrics, do ChatGPT and NMT systems perform better in English-to-Chinese (E-C) or Chinese-to-English (C-E) legal translation?

(3) What are the differences in the types of errors typically produced by ChatGPT and NMT systems?

II. Research Design

A. Source Texts

This study aims to carry out a comprehensive and systematic evaluation of the performance of NMT and ChatGPT in legal text translation. To ensure the validity and reliability of the research findings, the selection of source texts adheres to the following principles:

(1) comprehensiveness: Texts spanning various legal subfields, such as civil law, criminal law, commercial law, and administrative law, are included to ensure the research results are universally applicable and representative.

(2) timeliness: Only current and valid legal texts are chosen to reflect the real-world needs and challenges of legal translation accurately.

(3) diversity: Legal texts with different structures, levels of difficulty, and contextual nuances are selected to comprehensively assess the translation quality of legal texts by NMT and ChatGPT.

(4) authenticity: Selected laws and regulations are sourced from publicly available documents to facilitate peer review and verification of the research findings.

(5) referentiality: The chosen texts have official or authoritative translations available for reference, enabling the automated evaluation of the translation quality of NMT and ChatGPT.

Following the principles outlined above, 15 Chinese texts, each ranging from 500 to 550 characters, were selected from 14 different Chinese laws as the source texts (ST) for C-E translation (see Table 1). To ensure translation accuracy and authority, the English versions of these laws, provided by Chinalawinfo Database, were used as the standard reference translations for the target texts (TT). Similarly, to facilitate comparison with the ST, 15 corresponding English legal texts, also 500 to 550 words in length, were selected from the electronic version of Hong Kong Laws (available on elegislation.gov.hk) as the ST for E-C translation (see Table 1). The official Chinese versions of these texts, also obtained from the electronic version of the Hong Kong Laws, served as the TT standard reference translations.

TABLE 1
SOURCE TEXTS

	C-E		E-C	
No.	Law Titles in Chinese	Word Count	Law Titles in English	Word Count
1	《中华人民共和国宪法》	517	*The Basic Law of the Hong Kong Special Administrative Region of the People's Republic of China*	535
2	《中华人民共和国民法典》第五编 婚姻家庭	520	*Matrimonial Causes Ordinance*	549
3	《中华人民共和国刑法》	514	*Offences against the Person Ordinance*	541
4	《中华人民共和国企业破产法》	508	*Bankruptcy Ordinance*	504
5	《中华人民共和国民法典》第六编 继承	540	*Inheritance (Provision for Family and Dependants) Ordinance*	530
6	《中华人民共和国反洗钱法》	531	*Anti-Money Laundering and Counter-Terrorist Financing Ordinance*	530
7	《中华人民共和国著作权法》	521	*Copyright Ordinance*	509
8	《中华人民共和国证券法》	503	*Securities and Futures Ordinance*	525
9	《中华人民共和国劳动法》	511	*Employees' Compensation Ordinance*	515
10	《中华人民共和国刑事诉讼法》	501	*Costs in Criminal Cases Ordinance*	528
11	《中华人民共和国野生动物保护法》	525	*Wild Animals Protection Ordinance*	525
12	《中华人民共和国道路交通安全法》	500	*Road Traffic Ordinance*	529
13	《中华人民共和国印花税法》	536	*Stamp Duty Ordinance*	525
14	《中华人民共和国法律援助法》	533	*Legal Aid Ordinance*	528
15	《中华人民共和国教育法》	516	*Education Ordinance*	527
Total		7776		7877

B. Machine Translation Systems

This study selects ChatGPT and mainstream NMT systems as research subjects. ChatGPT is among the most widely used large language models in the world. According to Dai and Liu (2023), NMT is the primary focus of machine translation research. The four translation systems—Youdao Translate, Baidu Translate, Google Translate, and DeepL Translate—are all based on NMT technology. Of these, Google Translate is the most extensively studied NMT globally, while DeepL Translate claims to be "the best machine translation in the world." Youdao Translate and Baidu Translate are the most popular NMT systems in China.

C. Automated Evaluation Methods

In 2001, IBM introduced BLEU (Bilingual Evaluation Understudy) as a metric to evaluate the quality of machine translation. According to Wang and Wen (2010), BLEU has become a widely used index in the international machine translation evaluation system, with higher BLEU scores indicating better translation quality. This study will utilize the Translation Evaluation Tool provided on Shiyibao (https://www.shiyibao.com) to calculate BLEU scores for assessing translation quality.

D. Procedure

Firstly, I imported 30 source texts into Youdao Translate, Baidu Translate, Google Translate, and DeepL Translate, and then translated them using ChatGPT-4[1]. Secondly, the target texts generated by the NMT systems and ChatGPT-4 were copied into a Word document. Then, the BLEU scores of the target texts were calculated using the "Shiyibao - Translation Evaluation Tool". Finally, the BLEU values of the target texts were analyzed using SPSS 27 statistical software.

III. RESULTS

A. Comparison of C-E Translation Quality of Legal Texts Between ChatGPT and NMT

[1] The translations from the four NMT systems and ChatGPT-4 for this study were collected on December 8, 2023.

In this section, I will first compare the BLEU scores of ChatGPT with those of each NMT system for the C-E translation of legal texts, and then compare the BLEU scores of the NMT systems as a whole with those of ChatGPT.

(a). Comparison Between ChatGPT and Each NMT

Between ChatGPT and the four NMT systems, ChatGPT had the lowest average score and the highest standard deviation. This suggests that ChatGPT's C-E legal translations are comparatively lower in quality and less consistent. Among the four NMT systems, Youdao Translate achieved the highest average score with 70.07 points, followed by Google Translate with 66.52 points, while DeepL Translate and Baidu Translate had nearly identical scores. This analysis provides the descriptive statistics of BLEU scores for C-E translations, as presented in Table 2.

TABLE 2
DESCRIPTIVE STATISTICS OF OVERALL SCORES FOR C-E TRANSLATION

Translation Systems	Number of Scores	Mean	Standard Deviation	95% Confidence Interval for the Mean	
				Lower Limit	Upper Limit
ChatGPT	15	64.47	6.58	60.82	68.11
Youdao Translate	15	70.07	6.47	66.48	73.65
DeepL Translate	15	65.85	6.65	62.17	69.54
Baidu Translate	15	65.53	5.04	62.74	68.32
Google Translate	15	66.52	4.92	63.79	69.25
Total	75	66.49	6.13	65.08	67.90

To determine whether there are significant differences among the five translation systems in translating legal texts from Chinese to English, this study used SPSS 27 to test the normality of the BLEU scores for each system. The results indicated that the absolute values of kurtosis and skewness for all five datasets were less than 1.96. Additionally, the p-values for the Kolmogorov-Smirnov and Shapiro-Wilk tests were both greater than 0.05, suggesting that the BLEU scores in all five groups followed a roughly normal distribution.

Subsequently, a one-way ANOVA was conducted on the five groups of data to identify any significant differences among the translation systems. The analysis showed that there was no statistically significant difference among the five systems, with $p = 0.119$ (as seen in Table 3).

TABLE 3
VARIANCE TEST FOR OVERALL SCORES IN C-E TRANSLATION

	Sum of Squares	Degrees of Freedom	Mean Square	F	Significance
Between Groups	273.26	4	68.31	1.91	0.119
Within Groups	2507.35	70	35.82		
Total	2780.61	74			

Multiple comparison tests revealed a significant difference between ChatGPT and Youdao Translate among all translation systems. Within the NMT group, a significant difference was also observed between Baidu Translate and Youdao Translate, while the differences among the other translation systems did not reach statistical significance, as shown in Table 4.

TABLE 4
MULTIPLE COMPARISONS OF OVERALL SCORES FOR C-E TRANSLATION

	Translation Systems	Translation Systems	Mean Difference. (I-J)	Standard Error	Significance	95% Confidence Interval for the Mean	
						Lower Limit	Upper Limit
LSD	ChatGPT	Youdao Translate	-5.59933*	2.18538	0.013	-9.95794	-1.24072
		DeepL	-1.38467	2.18538	0.528	-5.74328	2.973945
		Baidu Translate	-1.06067	2.18538	0.629	-5.41928	3.297945
		Google Translate	-2.05333	2.18538	0.351	-6.41194	2.305278
	Youdao Translate	ChatGPT	5.59933*	2.18538	0.013	1.240722	9.957945
		DeepL	4.2146667	2.18538	0.058	-0.14394	8.573278
		Baidu Translate	4.53867*	2.18538	0.041	0.180055	8.897278
		Google Translate	3.546000	2.18538	0.109	-0.81261	7.904611
	DeepL Translate	ChatGPT	1.38467	2.18538	0.528	-2.97394	5.743278
		Youdao Translate	-4.21467	2.18538	0.058	-8.57328	0.143945
		Baidu Translate	0.32400	2.18538	0.883	-4.03461	4.682611
		Google Translate	-0.66867	2.18538	0.761	-5.02728	3.689945
	Baidu Translate	ChatGPT	1.060667	2.18538	0.629	-3.29794	5.419278
		Youdao Translate	-4.53867*	2.18538	0.041	-8.89728	-0.18006
		DeepL	-0.32400	2.18538	0.883	-4.68261	4.034611
		Google Translate	-0.99267	2.18538	0.651	-5.35128	3.365945
	Google Translate	ChatGPT	2.05333	2.18538	0.351	-2.30528	6.411945
		Youdao Translate	-3.54600	2.18538	0.109	-7.90461	0.812611
		DeepL	0.66867	2.18538	0.761	-3.68994	5.027278
		Baidu Translate	0.99267	2.18538	0.651	-3.36594	5.351278

*. The significance level for the mean difference is 0.05

In the C-E translation of legal texts, there is a significant difference between ChatGPT and Youdao Translate, but no significant difference between ChatGPT and the other three NMT systems. Among the four NMT systems, Baidu Translate has the lowest average score, showing a significant difference when compared to Youdao Translate, which has the highest average score. Additionally, DeepL Translate has a slightly higher average score than Baidu Translate.

(b). Overall Comparison Between ChatGPT and NMT

To compare the overall quality of C-E translations produced by ChatGPT and NMT systems, this study calculated the average BLEU scores for the four NMT systems across 15 legal texts. SPSS 27 was then used to assess the normality of the average BLEU scores in the 15 groups of C-E translations. The results showed that the absolute values of kurtosis and skewness were both less than 1.96, while the p-values for both the Kolmogorov-Smirnov and Shapiro-Wilk tests were greater than 0.05. This indicates that the BLEU scores for C-E translations by NMT systems follow a normal distribution.

Next, a one-way ANOVA was conducted to compare the mean BLEU scores for the C-E translation of legal texts by ChatGPT with the average scores for the four NMT systems. The analysis found no significant difference between the mean scores of ChatGPT and the NMT systems, with a p-value of 0.258 (as shown in Table 5).

Combining the mean value of ChatGPT (64.47) and the overall mean value of NMT (66.99), it is evident that the quality of C-E legal translations by ChatGPT is slightly lower than that of NMT, but there is no significant difference.

TABLE 5
VARIANCE TEST FOR OVERALL SCORES IN C-E TRANSLATION BY CHATGPT AND NMT

	Sum of Squares	Degrees of Freedom	Mean Square	F	Significance
Between Groups	47.855	1	47.855	1.334	0.258
Within Groups	1004.248	28	35.866		
Total	1052.103	29			

B. Comparison of E-C Translation Quality of Legal Texts Between ChatGPT and NMT

(a). Comparison Between ChatGPT and Each NMT

Based on the descriptive statistics for English-to-Chinese translation scores in Table 6, ChatGPT has the lowest

average score, while Youdao Translate has the highest. DeepL Translate comes next, followed by Baidu Translate and Google Translate.

TABLE 6
DESCRIPTIVE STATISTICS OF OVERALL SCORES FOR E-C TRANSLATION

Translation Systems	Number of Scores	Mean	Standard Deviation	95% Confidence Interval for the Mean	
				Lower Limit	Upper Limit
ChatGPT	15	29.55	11.67	23.09	36.01
Youdao Translate	15	43.43	13.73	35.83	51.03
DeepL Translate	15	42.88	15.32	34.39	51.36
Baidu Translate	15	41.12	10.92	35.07	47.16
Google Translate	15	37.66	13.59	30.13	45.18
Total	75	38.93	13.77	35.76	42.09

Since the absolute values of kurtosis and skewness for the scores of ChatGPT and the four NMT systems are all greater than 1.96, it can be inferred that the score distributions for ChatGPT and the four NMT systems are not normal. To determine whether there were significant differences among these systems, I employed the Kruskal-Wallis non-parametric test. The results indicated that there are indeed significant differences among the scores of the five systems, with $p < .001$ (see Table 7).

TABLE 7
KRUSKAL-WALLIS TEST STATISTICS FOR OVERALL E-C TRANSLATION SCORES [A,B]

	CE
Kruskal-Wallis H	23.050
Degrees of Freedom	4
Asymptotic Significance	<.001
A. Kruskal-Wallis Test	
B. Group Variable: Translation Systems	

Subsequently, I conducted a pairwise comparison of the five translation systems and found that at the significance level of .050, there were significant differences between ChatGPT and the four NMT systems, while there were no significant differences between the four NMT systems (see Table 8).

TABLE 8
PAIRWISE COMPARISON OF THE FIVE TRANSLATION SYSTEMS

Translation Systems	Test Statistics	Standard Error	Standard Inspection Statistics	Significance	Adj. Significance[a]
ChatGPT-Google Translate	-18.100	7.958	-2.274	0.023	0.229
ChatGPT-DeepL Translate	-28.867	7.958	-3.627	0.000	0.003
ChatGPT-Baidu Translate	-29.767	7.958	-3.740	0.000	0.002
ChatGPT-Youdao Translate	-33.100	7.958	-4.159	0.000	0.000
Google Translate-DeepL	10.767	7.958	1.353	0.176	1.000
Google Translate-Baidu Translate	11.667	7.958	1.466	0.143	1.000
Google Translate-Youdao Translate	15.000	7.958	1.885	0.059	0.594
DeepL Translate-Baidu Translate	-0.900	7.958	-0.113	0.910	1.000
DeepL Translate-Youdao Translate	4.233	7.958	0.532	0.595	1.000
Baidu Translate -Youdao Translate	3.333	7.958	0.419	0.675	1.000

Each row tested the null hypothesis that the distribution of translations by Translation System 1 is the same as that of Translation System 2.
The results indicate asymptotic significance (two-tailed test) with a significance level of .050.
a. The significance values have been adjusted for multiple tests using the Bonferroni correction method.

By analyzing the average scores presented in Table 6, it becomes evident that ChatGPT's proficiency in legal translation falls significantly behind that of the NMT systems. Among the four NMT systems evaluated, Youdao Translate demonstrates the highest standard in legal E-C translation, closely followed by DeepL Translate and Baidu Translate. In contrast, Google Translate exhibits the least satisfactory performance in legal E-C translation.

(b). Overall Comparison Between ChatGPT and NMT

To compare the quality of ChatGPT and NMT in E-C translation of legal texts, this study first computed the average BLEU scores for 15 legal documents across four NMT systems. Subsequently, SPSS 27 was employed to assess the normality of the BLEU scores for the 15 groups of E-C translations. The findings indicate that the absolute values of the mean kurtosis and skewness for the BLEU scores of the four NMT systems exceed 1.96, suggesting a lack of normal distribution in NMT English-to-Chinese translation scores. Next, this study utilized the Kruskal-Wallis non-parametric test to ascertain the presence of significant differences between the mean scores of ChatGPT and the NMT systems. Results indicate a significant difference between ChatGPT and NMT scores, with $p < .001$ (see Table 9).

In Table 6, ChatGPT's average score of 29.55 is markedly lower than the overall average score of 41.27 for NMT. Consequently, in E-C legal text translation, NMT demonstrates a notably higher quality of translation compared to ChatGPT.

TABLE 9
KRUSKAL-WALLIS TEST STATISTICS FOR OVERALL E-C TRANSLATION SCORES [A,B]

	EC
Kruskal-Wallis H	14.094
Degrees of Freedom	1
Asymptotic Significance	<.001

A. Kruskal-Wallis Test
B. Group Variable: Translation Systems

C. Comparison of Legal Text Translation Between E-C and C-E

The above analysis delves into the BLEU scores of ChatGPT and the four NMT systems in legal text translation between English and Chinese. In which direction of translation—E-C or C-E—do they excel when translating legal texts? Firstly, this study aggregates the scores of these five translation systems and conducts independent sample T-tests on their scores for both E-C and C-E translations, aiming to discern significant differences between them. The outcomes are detailed in Table 10 and Table 11.

As evident from Table 11, there are significant differences between ChatGPT and the four NMT systems in both E-C and C-E translations of legal texts (p [2-tailed] < .001). Moreover, the score for C-E translation ($M = 66.49$, $SD = 6.13$) significantly surpasses that of E-C translation ($M = 38.93$, $SD = 13.77$). Therefore, both ChatGPT and NMT exhibit superior performance in translating legal texts from Chinese to English.

TABLE 10
T-TEST GROUP STATISTICS

Translation Direction	Number of Scores	Mean	Standard Deviation	Mean Standard Error
C-E	75	66.49	6.13	0.71
E-C	75	38.93	13.77	1.59

TABLE 11
INDEPENDENT SAMPLES TEST

		Levene's Test for Equality of Variances		Equality of Means T-test						
		F	Significance	t	Degrees of Freedom	Significance (Two-tailed)	Mean Difference	Standard Error Difference	95% Confidence Interval for the Mean	
									Lower Limit	Upper Limit
BLEU 评分	Homogeneity of Variance	11.512	<.001	15.832	148	<.001	27.56133	1.740823	24.12125	31.00141
	Heterogeneity of Variance			15.832	102.208	<.001	27.56133	1.740823	24.10850	31.01416

IV. DISCUSSION

The preceding section presented a quantitative analysis to compare the overall translation quality of ChatGPT and NMT. In this section, a case study methodology is employed to closely examine the types of errors exhibited by each system. This study delineates the primary error categories in the translation of legal texts by ChatGPT and NMT as follows: terminology translation errors, grammatical and syntactic structure errors, and style and format errors. Subsequently, examples will be provided from both E-C and C-E translation perspectives.

A. Comparative Analysis of ChatGPT and NMT in C-E Translation of Legal Texts

This study randomly selected one paragraph from each of the 15 Chinese source texts. Subsequently, ChatGPT, Youdao Translate, DeepL Translate, Baidu Translate, and Google Translate were employed to translate these paragraphs. A comparative analysis of the translation results was then conducted.

Source Text

Example (1): 第二百三十四条 故意伤害他人身体的，处三年以下有期徒刑、拘役或者管制。

犯前款罪，致人重伤的，处三年以上十年以下有期徒刑；致人死亡或者以特别残忍手段致人重伤造成严重残疾的，处十年以上有期徒刑、无期徒刑或者死刑。本法另有规定的，依照规定。(https://flk.npc.gov.cn/detail2.html?ZmY4MDgxODE3OTZhNjM2YTAxNzk4MjJhMTk2NDBjOTI%3D)

Reference Translation

Example (1): Article 234. Whoever intentionally injures the person of another is to be sentenced to not more than three years of fixed-term imprisonment, limited incarceration, or probation.

Whoever commits the crime in the preceding paragraph and causes a person's serious injury is to be sentenced to not less than three years and not more than 10 years of fixed-term imprisonment; if he causes a person's death or causes a person's serious deformity by badly injuring him with particularly ruthless means, he is to be sentenced to not less than 10 years of fixed-term imprisonment, life imprisonment, or death. Where this Law has other stipulations, matters are to be handled in accordance with such stipulations. (https://www.pkulaw.com/en_law/39c1b78830b970eabdfb.html)

ChatGPT-4 Translation

Example (1): Article 234: Whoever intentionally injures another person's body shall be sentenced to fixed-term imprisonment of not more than three years, criminal detention, or public surveillance.

Whoever commits the crime mentioned in the preceding paragraph and causes serious injury to a person shall be sentenced to fixed-term imprisonment of not less than three years but not more than ten years; if he causes the death of a person or causes serious disability to a person by particularly cruel means, he shall be sentenced to fixed-term imprisonment of not less than ten years, life imprisonment, or death. Where this law provides otherwise, such provisions shall apply.

Youdao Translate Translation

Example (1): Article 234 Whoever intentionally injures another person shall be sentenced to fixed-term imprisonment of not more than three years, criminal detention or public surveillance.

Whoever commits the crime mentioned in the preceding paragraph, thereby causing serious injury to another person, shall be sentenced to fixed-term imprisonment of not less than three years but not more than 10 years; If he causes death to another person or causes serious injury to another person by especially cruel means, resulting in serious disability, he shall be sentenced to fixed-term imprisonment of not less than 10 years, life imprisonment or death. Where this Law provides otherwise, such provisions shall prevail.

DeepL Translate Translation

Example (1): Article 234 Whoever intentionally inflicts bodily harm on another person shall be sentenced to fixed-term imprisonment of not more than three years, criminal detention or control.

Whoever commits the crime in the preceding paragraph and causes serious injury to a person shall be sentenced to fixed-term imprisonment of not less than three years and not more than ten years; whoever causes death or causes serious injury to a person by particularly cruel means resulting in serious disability shall be sentenced to fixed-term imprisonment of not less than ten years, life imprisonment or the death penalty. Where this Law provides otherwise, it shall do so in accordance with the provisions.

Baidu Translate Translation

Example (1): Article 234 Whoever intentionally harms another person's body shall be sentenced to fixed-term imprisonment of not more than three years, criminal detention, or public surveillance.

Whoever commits the crime mentioned in the preceding paragraph and causes serious injury to a person shall be sentenced to fixed-term imprisonment of not less than three years but not more than ten years; Those who cause death or serious injury to a person through particularly cruel means, resulting in serious disability, shall be sentenced to fixed-term imprisonment of not less than ten years, life imprisonment, or death. If there are other provisions in this Law, they shall prevail.

Google Translate Translation

Example (1): Article 234 Anyone who intentionally harms the body of another person shall be sentenced to fixed-term imprisonment of not more than three years, criminal detention or public surveillance.

Whoever commits the crime in the preceding paragraph and causes serious injury to another person shall be sentenced to fixed-term imprisonment of not less than three years but not more than ten years; whoever causes death or serious injury and severe disability by particularly cruel means shall be sentenced to fixed-term imprisonment of not less than ten years but not more than ten years, or life imprisonment or death. If this law provides otherwise, the provisions shall prevail.

Using manual translation as a reference, the translations generated by ChatGPT and the four NMT systems were compared against it.

Overall, both ChatGPT and the four NMT systems manage to convey the fundamental information of Example 1 with relatively accurate expressions. However, several issues arise, primarily stemming from terminology translation errors, grammatical and syntactic structure errors, as well as style and format errors.

The first issue concerns the accuracy of terminology translation. Example (1) involves several legal terms, such as "有期徒刑、拘役、管制、无期徒刑、死刑". ChatGPT and the four NMT systems generally provide accurate translations

for these terms. Both ChatGPT and the four NMT systems render “有期徒刑” and “无期徒刑” as “fixed-term imprisonment” and “life imprisonment” respectively, which aligns with the reference translation and is correct. Similarly, ChatGPT and the other NMT systems consistently translate “死刑” as “death”, which is consistent with the reference translation. However, DeepL’s translation of “死刑” as “the death penalty” appears somewhat cumbersome.

In addition, there are discrepancies in the translation of “拘役”. ChatGPT and the four NMT systems translate “拘役” as “criminal detention”, which differs from “limited incarceration” in the reference translation. Yet, this translation is also accurate. DeepL translates “管制” as “control”, which lacks precision and may lead to ambiguity. Both ChatGPT and other NMT systems translate “管制” as “public surveillance”, which differs from the reference translation. However, from the perspective of legal terminology, this translation is also accurate. Overall, in terms of terminology translation, both ChatGPT and NMT perform comparably, exhibiting very high accuracy levels that make it difficult to distinguish clear differences in their performance.

The second problem concerns the translation of grammar and sentence structures. In the first sentence of Example (1), Google Translate uses the sentence “Anyone who intentionally harms the body of another person shall be sentenced to,” whereas other NMT systems and ChatGPT utilize the sentence pattern “Whoever intentionally... shall be sentenced to.” Although the two sentence patterns differ slightly, they both conform to the standards of legal English expression.

However, when translating “十年以上有期徒刑” in the second sentence of Example (1), Google Translate made an obvious error by stating “not less than ten years but not more than ten years,” which is contradictory and may lead to ambiguity, confusing readers. As for the translation of “致人死亡或者以特别残忍手段致人重伤造成严重残疾的,” ChatGPT’s rendition, “if he causes the death of a person or causes serious disability to a person by particularly cruel means,” is concise and clear. However, there is ambiguity in the NMT’s translation of this sentence. For example, Youdao Translate renders it as “if he causes death to another person or causes serious injury to another person by especially cruel means, resulting in serious disability.” This sentence consists of two conditions: “if he causes death to another person” and “or causes serious injury to another person by especially cruel means, resulting in serious disability.” These two conditions specify two possible outcomes: “death” or “serious disability”. Understanding this sentence requires considering the relationship between the two conditions. Specifically, if someone causes the death of another person (condition 1), then the “or” in condition 2 can be deemed not to apply because the first condition has already been met. But if someone does not cause death, but rather causes someone else to be severely disabled by particularly cruel means (condition 2), then “serious disability” will eventually result. Although this sentence is understandable, it may in some cases require careful thought by the reader to ensure that the relationship between the conditions is properly understood. This complexity can lead to some potential ambiguities, especially in legal texts, where the law requires clear and accurate expression.

The third problem pertains to errors in style and format. Both ChatGPT and NMT exhibit no apparent formatting errors, and their translation structures align with the original texts, adhering to the typical format of legal documents. They predominantly employ standard legal English terms and sentence patterns to ensure accuracy and legal precision. However, there are instances where translations deviate from legal English norms. For instance, DeepL’s phrase “intentionally inflicts bodily harm” appears slightly blunt, and Baidu Translate’s use of “those who...” is not very common in legal English.

Overall, while the translations of legal texts by NMT and ChatGPT are relatively accurate in terms of style and format, there is still room for improvement, particularly in maintaining sentence consistency and avoiding obvious translation errors. By incorporating these enhancements, the translation quality can be elevated, making legal texts more precise and understandable.

B. Comparative Analysis of ChatGPT and NMT in E-C Translation of Legal Texts

This study randomly selected one text from each of the 15 English source texts. Subsequently, ChatGPT, Youdao Translate, DeepL Translate, Baidu Translate, and Google Translate were employed to translate these texts. A comparative analysis of the translation results was then conducted.

Source Text

Example (2): 11. Destroying or damaging building with intent to murder

Any person who, by the explosion of gunpowder or any other explosive substance, destroys or damages any building with intent to commit murder shall be guilty of an offence triable upon indictment, and shall be liable to imprisonment for life.

(Amended 30 of 1911 ss. 2, 4 & 5; 50 of 1991 s. 4)

[cf. 1861 c. 100 s. 12 U.K.] (https://www.elegislation.gov.hk/hk/cap212)

Reference Translation

Example (2) : 11. 意图谋杀而摧毁或破坏建筑物

任何人意图谋杀而藉火药或任何其他爆炸品的爆炸以摧毁或破坏建筑物，即属犯可循公诉程序审讯的罪行，可处终身监禁。

(由 1911 年第 30 号第 2、4 及 5 条修订；由 1991 年第 50 号第 4 条修订)

[比照 1861 c. 100 s. 11 U.K.] (https://www.elegislation.gov.hk/hk/cap212!sc?INDEX_CS=N)

ChatGPT-4 Translation
Example (2) : 11. 故意破坏或损坏建筑物以谋杀
任何人通过爆炸火药或任何其他爆炸物破坏或损坏任何建筑物，意图谋杀，应被视为犯有可被起诉的罪行，并可能面临终身监禁。
(修订于 1911 年第 30 号法案第 2、4 和 5 条；1991 年第 50 号法案第 4 条)
[参考 1861 年英国法案第 100 号第 12 条]
Youdao Translate Translation
Example (2) : 11. 蓄意毁坏或破坏建筑物的
任何人因爆炸火药或任何其他爆炸性物质而意图谋杀而毁坏或损坏任何建筑物，即属犯罪，可经检控审讯，并可处终身监禁。
(由 1911 年第 30 号第 2、4 及 5 条修订;1991 年第 50 条第 4 款)
[参见 1861 年，约 100 年，英国 12 年]
DeepL Translate Translation
Example (2) : 11. 意图谋杀而摧毁或损坏建筑物
任何人意图谋杀而藉爆炸火药或任何其他爆炸物质毁坏或损坏任何建筑物，即属犯可循公诉程序审讯的罪行，可处终身监禁。
(由 1911 年第 30 号第 2、4 及 5 条修订；由 1991 年第 50 号第 4 条修订）。
[比 照 1861 c. 100 s. 12 U.K.] 12.
Baidu Translate Translation
Example (2) : 11. 意图谋杀而毁坏或损毁建筑物
任何人藉火药或任何其他爆炸性物质的爆炸，意图谋杀而毁坏或损坏任何建筑物，即属犯可循公诉程序审讯的罪行，可处终身监禁。
（由 1911 年第 30 号第 2、4 及 5 条修订；由 1991 年第 50 号第 4 条修订）
[比照 1861 c.100 s.12 英国]
Google Translate Translation
Example (2) : 11. 意图谋杀而毁坏或损坏建筑物
任何人通过火药或任何其他爆炸性物质爆炸，摧毁或损坏任何建筑物，意图实施谋杀，均属犯罪，可经公诉审理，并可判处终身监禁。
(由 1911 年第 30 号第 2、4 及 5 条修订；由 1991 年第 50 号第 4 条修订)
[参见。1861 年 100 秒。12 英国]

The first problem concerns the accuracy of terminology translation. Example (2) involves several legal terms, including “with intent to murder”, “be guilty of an offence triable upon indictment”, and “imprisonment for life”. Both ChatGPT and NMT accurately translate “imprisonment for life” as “终身监禁”. However, NMT provides a better translation for “with intent to murder” and “be guilty of an offence triable upon indictment”. ChatGPT’s translation of “with intent to murder” as “以谋杀” is overly simplified and fails to capture the legal intent conveyed in the original text. The original expression clearly indicates that the intent behind the act was to commit murder, whereas ChatGPT’s translation is relatively vague and may not sufficiently convey the legal rigor of the original text.

Additionally, ChatGPT translates “be guilty of an offence triable upon indictment” as “犯有可被起诉的罪行”, which conveys the general meaning but overlooks the crucial legal step of “indictment” mentioned in the original text. This translation lacks the necessary legal precision and may introduce ambiguity into the legal text by failing to clearly express the trial process.

In general, NMT provides more accurate and precise translations of legal terms between Chinese and English, which can better meet the requirements for translating legal texts into English. The translation of legal texts requires a deep understanding of legal terms and procedures to ensure the accuracy and precision of legal translation.

The second issue pertains to grammar and sentence pattern translation. Overall, NMT outperforms ChatGPT in terms of grammatical accuracy and sentence structure. Example (2) illustrates that ChatGPT sometimes generates expressions that are informal or insufficiently accurate when handling professional legal texts. In contrast, NMT can more precisely convey the meaning of the source text using standardized expressions. For example, DeepL translates “shall be guilty of an offence triable upon indictment, and shall be liable to imprisonment for life” as “即属犯可循公诉程序审讯的罪行，可处终身监禁”. The sentence structure is clear and standardized, directly conveying the meaning of the original text.

In general, NMT excels in grammatical structure, accuracy in literal translation, and formal expression. When tasked with highly professional and normative legal texts, NMT may better ensure the accuracy and professionalism of the translation results.

The third issue pertains to style and format. For instance, when translating “(Amended 30 of 1911 ss. 2, 4 & 5; 50 of 1991 s. 4),” ChatGPT renders it as “修订于 1911 年第 30 号法案第 2、4 和 5 条;1991 年第 50 号法案第 4 条.” However, compared to ChatGPT, NMT’s translation appears more standardized and better aligns with standard Chinese usage.

Consequently, NMT outperforms ChatGPT in terms of handling style and format.

In conclusion, NMT not only provides more accurate term translation but also excels in grammatical structure, literal translation accuracy, and formal expression in E-C legal translation.

V. Conclusions

This study compared the performance of ChatGPT and four mainstream NMT systems in translating legal texts between English and Chinese using both quantitative and qualitative analyses. The results indicate that both ChatGPT and NMT systems meet the basic requirements for Chinese-to-English legal translation. While ChatGPT performs slightly worse than the NMT systems, the difference between them is not statistically significant. However, in the task of translating legal texts from English to Chinese, both systems fail to meet the required standards, with the NMT systems performing relatively better. Overall, ChatGPT and NMT systems demonstrate strong capabilities in translating legal texts from Chinese to English. However, ChatGPT exhibits less accuracy and reliability than NMT when translating legal texts between English and Chinese. Additionally, while both systems exhibit similar types of translation errors, ChatGPT tends to have more frequent and severe errors.

This study provides valuable insights for readers seeking guidance on selecting legal translation tools for English-to-Chinese and Chinese-to-English translations. Furthermore, it highlights the persistent challenges encountered by machine translation systems, especially in specialized fields such as legal translation. This emphasizes the necessity for continuous optimization to attain high standards of translation quality.

References

[1] Cady, L., Tsou, B. K., & Lee, J. S. (2023). Comparing Chinese-English MT Performance Involving ChatGPT and MT Providers and the Efficacy of AI mediated Post-Editing. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track* (pp. 205-216).

[2] Feng, Z., & Zhang, D. (2022). Machine Translation and Human Translation Boost Each Other. *Journal of Foreign Languages, (06)*, 77-87.

[3] Grimm, D. R., Lee, Y. J., Hu, K., Liu, L., Garcia, O., Balakrishnan, K., & Ayoub, N. F. (2024). The utility of ChatGPT as a generative medical translator. *European Archives of Oto-Rhino-Laryngology*, 1-5.

[4] Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., ... & Awadalla, H. H. (2023). *How good are GPT models at machine translation? A comprehensive evaluation*. arXiv preprint arXiv:2302.09210.

[5] Jiao, W., Wang, W., Huang, J. T., Wang, X., Shi, S., & Tu, Z. (2023). *Is ChatGPT a good translator? Yes with GPT-4 as the engine*. arXiv preprint arXiv:2301.08745.

[6] Khoshafah, F. (2023). *ChatGPT for Arabic-English translation: Evaluating the accuracy*.

[7] Lee, T. K. (2023). Artificial intelligence and posthumanist translation: ChatGPT versus the translator. *Applied Linguistics Review*. https://doi.org/10.1515/applirev-2023-0122.

[8] Li, F. (2021). A Comparative Study on the Quality of English-Chinese and Chinese-English Translations in Neural Network-Based Online Machine Translation Systems. *Shanghai Journal of Translators, (04)*, 46-52.

[9] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318).

[10] Retrieved October 2023, from *https://flk.npc.gov.cn/detail2.html?ZmY4MDgxODE3OTZhNjM2YTAxNzk4MjJhMTk2NDBjOTI%3D*

[11] Retrieved October 2023, from *https://www.elegislation.gov.hk/hk/cap212*

[12] Retrieved October 2023, from *https://www.elegislation.gov.hk/hk/cap212!sc?INDEX_CS=N*

[13] Retrieved October 2023, from *https://www.pkulaw.com/en_law/39c1b78830b970eabdfb.html*

[14] Sanz-Valdivieso, L., & López-Arroyo, B. (2023). Google Translate vs. ChatGPT: Can non-language professionals trust them for specialized translation? In *International Conference Human-informed Translation and Interpreting Technology (HiT-IT 2023)* (pp. 97-107).

[15] Wang, J., & Wen, Q. (2010). A Review of Domestic and International Automated Machine Scoring Systems: Implications for Automated Translation Scoring Systems for Chinese Students. *Foreign Language World, (1)*, 75-81+91.

[16] Wang, Z., & Mao, C. (2023). Assessing and Improving the Quality of ChatGPT Translations: A Case Study of Chinese-to-English Translations of Ceramic-Related Texts. *Shangdong Ceramics, (04)*, 20-27.

[17] Yang, F. (2023). Reflections and Inspirations from ChatGPT: A Case Study of Vietnamese Legal Translation. *Chinese Science & Technology Translators Journal, (03)*, 27-30+4.

[18] Yu, L. (2024). ChatGPT Lexical diversity and syntactic complexity in ChatGPT translation. *Foreign Language Teaching and Research, (02)*, 297-307+321. doi:10.19923/j.cnki.fltr.2024.02.005.

[19] Zhao, R., Huang, Y., Ma W., Dong, W., Xian, G., &Sun, T. (2023). Insights and Reflections of the Impact of ChatGPT on Intelligent Knowledge Services in Libraries. *Journal of Library and Information Science in Agriculture, (01)*, 29-38.

[20] Zhou, C., & Liu, Z. (2022). Machine Translation of Ancient Chinese Text Based on Transformer of Semantic Information Sharing. *Technology Intelligence Engineering, (06)*, 114-127.

Lijie Ding is currently pursuing her Master's degree in linguistics at the School of Foreign Languages, Southwest University of Political Science and Law in Chongqing, China. She has received professional training in English-Chinese translation and corpus-based research. Her research interests include English-Chinese legal translation and corpus linguistics.