Journal of Artificial Intelligence Practice (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2025.080310
ISSN 2371-8412 Vol. 8 Num. 3

An Empirical Analysis of Neural Network Machine Translation and Human Translation

Xinyue Li

School of Interpreting and Translation Studies, Guangdong University of Foreign Studies, Guangzhou, Guangdong, 510420, China

Keywords: Machine Translation; Human Translation; Quantitative Analysis; Sentiment Analysis

Abstract: This study explores the performance and sentiment analysis differences between machine translation and human translation in various fields from a quantitative perspective. The findings reveal that: 1. Human translation performs better than machine translation across all fields, but the gap between machine and human translation is smaller in daily language. 2. In highly creative fields such as literature, there is still a significant gap between machine translation and human translation. 3. In specialized fields like economy, trade, and business, which have a high volume of professional vocabulary, the accuracy of machine translation may decrease if the text database is not updated promptly. 4. Machine translation is in urgent need of improvement in aspects such as vocabulary richness, expression flexibility, complex sentence recognition, sentiment analysis and expression, and translation methods. Overall, this study not only deepens the path of empirical translation research but also offers insights for the study of translation technology.

1. Introduction

With machine translation growing more widespread, arguments like "technological substitution theory" and "technological threat theory" have often confused translation researchers. While many scholars have explored the differences between machine and human translation[1-4], most rely on qualitative approaches, lack quantitative empirical studies, and focus narrowly on common fields like politics and news, neglecting analysis of other vertical domains.

This study supplements prior research by comparing the translation accuracy of human and machine translation across different domains and innovatively incorporating sentiment analysis. It selects source texts from various fields for training neural network models, conducts horizontal comparisons among mainstream translation software, the neural network machine translation model developed in this study, and human translation, and aims to objectively evaluate both translation modes while promoting future research on human-machine collaborative translation.

2. Model Training Process

Instance-based machine translation completes the automatic translation process through the

following three steps.1. The system searches for the source language segments to be translated within the source language corpus and stores all the related segments found; 2. The system searches for equivalent target language translations in parallel texts; 3. The target language segments found are reorganized into the target language discourse [5]. After processing the data from the corpus, this study begins to construct a translation model based on the LSTM neural network.

For example, if the source corpus is a Chinese sentence, the source and target corpora will be matched word by word via segmentation. Then, the target corpus will be cut into "I/ went / to /the library/ to/ read /books/.", and each word will be vectorized by generating a decoder. Each word is then fed into the neural network model on the right hand side, where each word corresponds to a layer of the neural network, i.e., a neural layer. After the training data of the neural network (vectorized vocabulary in English) and the target data (vectorized vocabulary in Chinese), the weight relationship between the source language vocabulary and the target language vocabulary can be fitted, which forms a model, and the strength of the model depends on how much vocabulary is available, the larger the vocabulary is, the better the model is fitted and the more accurate the translation is. When a new vocabulary is input, the neural network multiplies the encoding of the vocabulary by the weights of each layer to derive the probability of the target language vocabulary, which is finally translated into the target language vocabulary by the decoder. In order to reduce the amount of computation and fit the habit of human translation, this study introduces an attention mechanism when calculating the vocabulary weights, so that the weights of each source language vocabulary are biased towards the target language vocabulary, thus improving the translation accuracy.

3. Model Settings and Evaluation Criteria

3.1 Evaluation criteria for the model

In the process of LSTM model training, the cross-entropy function is set as the loss function, and the cross entropy is calculated as follows:$H(p,q) = -\sum_i p(i) log q(i)$, where denotes the true value, denotes the predicted value, and denotes the cross-entropy loss. For example, the true target label is [0,0,1] and the predicted value is [0.26,0.24,0.5], then the cross-entropy loss function $\text{Loss} = -0 * \log(0.26) - 0 * \log(0.24) - 1 * \log(0.5) \approx 0.301$. In Python, as epochs are iterated and the weights of each neural layer are updated, the loss function of the model will decrease, see Figure 1.

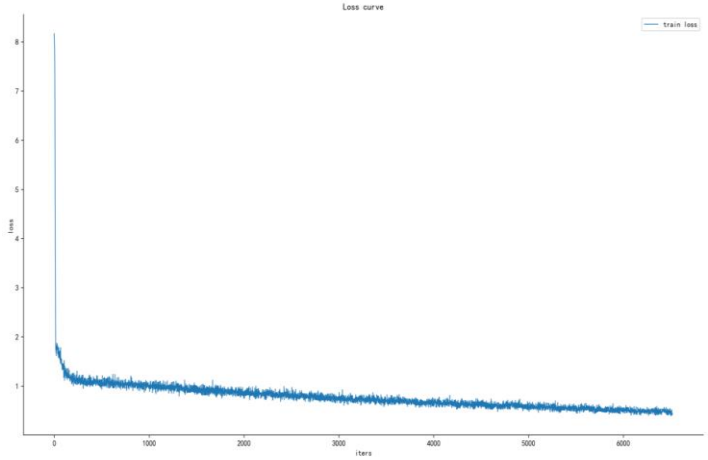

Figure 1: Loss function plot

As shown in Figure 1, when training is conducted only once, the loss function Loss value is relatively high, at 8.17, indicating a significant difference between the predicted values and the true

values. By the time training reaches the fifth round, Loss has already fallen to 0.85, gradually increasing the model's fitting ability. The trend characteristics of the loss function can also be observed from the loss function (Loss curve) graph. In the first round of training, the loss function value is positioned high, but after 200 iterations, it drops to around 1 and continues to fluctuate around 1, indicating that the model's training is becoming more stable.

3.2 Accuracy evaluation of model predictions

The BLEU metric can be used to evaluate the translation performance of training models. The BLEU value ranges from 0 to 1, with values closer to 1 indicating higher translation quality. This study introduces an improved evaluation criterion based on the original BLEU, called n-gram precision.

$$P_n = \frac{\sum_{i \in n\_gram} \min(h_i(C), max_{j \in m} h_i(S_i))}{\sum_{i \in n\_gram} h_i(C)} \quad (1)$$

Where is the number of times $n\_gram\ i$ appears inside C; $h_i(S_i)$ denotes the number of times i appears in the i^{th} reference translation.

4. Empirical Analysis

This section examines human-machine translation differences. It uses texts from three domains: daily language, business and trade, and literature to construct neural network models, then randomly selects 100 sentences per domain for translation. To explore deep neural network machine translation's domain-specific applicability, it horizontally compares with other neural network translation tools (DeepL, Youdao, Tencent, and Baidu Translate).

4.1 Translation in the field of daily language

The following is a statistical description of the BLEU evaluation values for each software and the neural network translation model constructed in this study in the domain of daily language.

Table 1: Statistical description and BLEU correlation coefficients: machine vs. human translation (daily language domain)

Panel A Statistical Description						
Daily words	LSTM network	DeepL	Baidu Translation	Tencent Translation	Youdao Translation	Human Translation
Mean	0.697	0.695	0.707	0.701	0.711	0.848
Median	0.695	0.693	0.714	0.694	0.706	0.849
Max	0.934	0.921	0.871	0.905	0.946	0.890
Min	0.474	0.459	0.423	0.512	0.569	0.806
Std.Dev	0.084	0.074	0.088	0.079	0.083	0.020
Panel B Correlation coefficient						
Neural network	1.000					
DeepL	-0.024	1.000				
Baidu Translation	-0.008	-0.001	1.000			
Tencent Translation	-0.067	0.117	-0.209***	1.000		
Youdao translation	0.027	-0.042	0.194**	-0.058	1.000	
Human translation	-0.077	0.024	-0.123	0.172	0.135	1.000

Note: ***, **,* denote significance at the 1%, 5% and 10% levels, respectively.

As shown in Table1, BLEU values are calculated for individual machine translations separately, where n-gram is taken as 4. It can be seen that the average BLEU evaluation index of individual

machine translations is around 0.7. Among the individual translations, the highest accuracy is found in the Youdao translation (Max=0.946) and the lowest in the Baidu translation (Max=0.871). From the average, minimum, and maximum values of BLEU, among the selected samples, the accuracy of the Youdao translation is higher. From the constructed LSTM neural network translation, its performance is moderate, only slightly better than DeepL. This may be attributed to the smaller training sample size of this study's neural network, which cannot achieve the effectiveness of traditional machine translation software trained on large samples. To compare the differences and similarities between human and machine translations, this study invites five professional English translators to conduct human translations. Their sentences were compared to reference texts using the BLEU value calculation, with an average BLEU value of about 0.80, indicating a higher translation accuracy than the average accuracy of machine translations. The study also discusses the correlations between various machine translations. Panel B shows that Baidu Translation and Tencent Translation are significantly negatively correlated. This may be due to the difference in translation accuracy between Baidu and Tencent when translating sentences in different domains, with Tencent Translation possibly misinterpreting some Chinese words. Baidu and Youdao show a positive correlation in translation accuracy, indicating a high level of synchronicity in their Chinese translation capabilities.

To explore the differences in accuracy between machine and human translations, this study employs a bivariate t-test to examine the translations of daily language by human translators and various machine translations. The results are shown in Table 2:

Table 2: T-test difference analysis: machine vs. human translation (daily language domain)

Daily words	Human VS. LSTM networks	Human VS. DeepL	Human VS. Baidu	Human VS. Tencent	Human VS. Youdao
Human-Machine	0.151***	0.153***	0.141***	0.147***	0.137***
P-value	0.000	0.000	0.000	0.000	0.000
T-value	17.199	20.108	15.274	18.828	16.619

Note: ***,**,* indicate significance at the 1%, 5% and 10% levels, respectively.

As shown in Table 2, a two-sample t-test compares the average BLEU value of human translation with that of the LSTM model, DeepL, Baidu, Tencent, and Youdao. The null hypothesis: $\alpha 1=\alpha 2\ldots=0$, indicates that there is no significant difference between the BLEU values of the two variables. If the results of the t-test reject the null hypothesis, then there is a significant difference between the BLEU values of the two variables. According to above results, there is a statistically significant difference between machine and human translations. The difference between human translation and neural network translation is 0.151, with a P-value of 0.000 and a T-value of 17.199, indicating a significant difference in translation accuracy between human and neural network translations. Other translation software yields similar results: human translation is more accurate and has distinct advantages.

4.2 Translation in the field of trade and economy

The following is a statistical description of the BLEU evaluation values of each machine translation software and the neural network translation model constructed in this study in the field of economy and trade.

From Table 3, it can be observed that the average BLEU evaluation values for various machine translations are around 0.5. In contrast, human translation achieves a BLEU value of 0.735, which is approximately 47% higher than machine translation. Among individual sentences in machine translation, the neural network machine translation achieves the highest accuracy (Max=0.997), while Youdao Translation has the lowest (Max=0.987). From the average, minimum, and maximum

values of BLEU, human translation exhibits higher translation accuracy in these 100 example sentences, with experimental results being effective within the sample. Regarding the LSTM neural network translation constructed in this study, its performance in the field of economy and trade is moderate, slightly better than Youdao Translation. This might be attributed to the extensive use of specialized terms in the field of economy and trade, resulting in lower translation accuracy outside the sample. In Panel B, DeepL is significantly negatively correlated with LSTM, while it is significantly positively correlated with Baidu Translation.

Table 3: Statistical description and BLEU correlation coefficients: machine vs. human translation (economic and trade fields)

Panel A Statistical Description						
Economic trade	LSTM network	DeepL	Baidu Translation	Tencent Translation	Youdao Translation	Human Translation
Mean	0.483	0.529	0.502	0.526	0.458	0.735
Median	0.471	0.515	0.490	0.521	0.468	0.738
Max	0.997	0.995	0.996	0.991	0.987	0.994
Min	0.004	0.003	0.016	0.005	0.005	0.466
Std.Dev	0.273	0.283	0.301	0.285	0.278	0.099
Panel B Correlation coefficient						
Neural network	1					
DeepL	-0.236**	1				
Baidu Translation	-0.017	0.233**	1			
Tencent Translation	-0.101	0.119	0.067	1		
Youdao Translation	-0.059	0.154	-0.097	0.106	1	
Human Translation	-0.082	0.023	0.095	0.023	0.028	1

Note: ***, **,* indicate significance at the 1%, 5% and 10% levels, respectively.

Table 4: T-test difference analysis: machine vs. human translation (economic and trade fields)

	Human VS. LSTM networks	Human VS. DeepL	Human VS. Baidu	Human VS. Tencent	Human VS. Youdao
Human-Machine	0.25***	0.21***	0.23***	0.21***	0.28***
P-value	0.00	0.00	0.00	0.00	0.00
T-value	8.452	6.897	7.560	6.991	9.476

Note: ***, **,* denote significance at the 1%, 5% and 10% levels, respectively.

As shown in Table 4, the average BLEU value for human translation is approximately 0.735, indicating a high level of translation accuracy. Similarly, a two-sample t-test was conducted in the study. It can be seen that there is a significant statistical difference between machine translation and human translation. The P-value for Human Translation vs. Neural Network is 0.00, with a T-value of 8.452, indicating a significant difference in translation accuracy between human and neural network translations. Similar results are observed for the other comparisons, indicating that in economic and trade translation, human translation is statistically more accurate than machine translation, highlighting its unique advantages.

4.3 Translation in the field of literary works

The following is a statistical description of the BLEU scores for each machine translation software and the neural network translation model constructed in this study in the field of literature.

Table 5 still shows human translation's average BLEU value exceeds that of machine translation and neural networks. While machine translation outperforms human translation in some individual statements, neural networks, DeepL, and Youdao achieve BLEU scores over 0.99, indicating such tools can attain high accuracy for specific texts. Overall, however, human translation in the selected samples is more accurate. With a standard deviation (Std. Dev) of only 0.0878, its accuracy shows

low deviation and small variance across individual statements, whereas machine translation has a large deviation. The LSTM neural network translation constructed here performs better in literary translation, ranking only below human translation and DeepL. In this domain, only DeepL correlates significantly with the LSTM model.

Table 5: Statistical description and BLEU correlation coefficients: machine vs. human translation (literary field)

Panel A Statistical Description						
Literary	LSTM network	DeepL	Baidu Translation	Tencent Translation	Youdao Translation	Human Translation
Mean	0.4973	0.5744	0.4727	0.4538	0.4879	0.7164
Median	0.5030	0.6287	0.4539	0.4194	0.5114	0.7152
Max	0.9967	0.9993	0.9888	0.9754	0.9983	0.9201
Min	0.0018	0.0116	0.0007	0.0004	0.0146	0.4930
Std. Dev	0.3267	0.2641	0.2776	0.2921	0.2885	0.0878
Panel B Correlation coefficient						
Neural network	1					
DeepL	-0.313***	1				
Baidu Translation	-0.034	-0.104	1			
Tencent Translation	0.094	0.002	-0.027	1		
Youdao Translation	-0.151	0.118	0.005	0.055	1	
Human Translation	0.051	-0.179*	-0.135	0.05	-0.067	1

Note: ***, **, * indicate significance at the 1%, 5% and 10% levels, respectively.

This study also uses a two-sample t-test to verify whether human translation significantly outperforms machine translation.

Table 6: T-test difference analysis: machine vs. human translation (literary field)

literary	Human VS. LSTM networks	Human VS. DeepL	Human VS. Baidu	Human VS. Tencent	Human VS. Youdao
Human-Machine	0.219***	0.142***	0.244***	0.263***	0.229***
P-value	0.000	0.000	0.000	0.000	0.000
T-value	6.562	4.850	8.067	8.733	7.441

Note: ***, **,* indicate significance at the 1%, 5% and 10% levels, respectively.

Table 6 reports the t-test for human-machine translation accuracy differences in the literary domain, calculated by subtracting machine translation's average BLEU value from human translation's. From the test results, the accuracy of human translation is statistically significantly better than the other machine translations, in which the human translation is 0.263 (t-value = 8.733) more accurate than Tencent Translation, which rejects the null hypothesis at 1% level that the accuracy of the human translation is significantly higher than the Tencent Translation, and that the human translation is optimal in the comparison with the rest of the software.

5. Further Discussion: Corpus Analysis

To further compare human and machine translation performance and quality in specific fields, this study selects materials from business negotiations, literature, and speeches for comprehensive analysis. The source language text consists of 4,251 English words, while the target language includes five Chinese texts, with human translations containing 7,134 words, DeepL translations 7,114 words, Tencent translations 7,082 words, Youdao translations 6,916 words, and Baidu translations 6,825 words. Due to the limited training corpus for the LSTM model, it was excluded from the following research, focusing only on a horizontal comparison between several translation software and human translations.

This study references Chinese quantitative features for text clustering from reference [6-8],

selecting word length, sentence length, type-token ratio, adverb ratio, and pronoun ratio for analysis. Conjunctions and adjectives are also included, given their roles in logical rigor and lexical richness. Subsequently, this study employs Jieba segmentation library in Python for text segmentation and SnowNLP for part-of-speech tagging, calculating the proportions of nouns, adverbs, adjectives, and other linguistic structures in the texts to determine the distribution of 12 linguistic structures. The formulas are as follows: word length = number of characters / number of words; sentence length = number of characters / number of sentences; type-token ratio = number of words / number of word types; adverb ratio = number of adverbs / number of words; noun ratio = number of nouns / number of words; verb ratio = number of verbs / number of words; pronoun ratio = number of pronouns / number of words; conjunction ratio = number of conjunctions / number of words; adjective ratio = number of adjectives / number of words; declarative sentence ratio = number of declarative sentences / total number of sentences; interrogative sentence ratio = number of interrogative sentences / total number of sentences; exclamatory sentence ratio = number of exclamatory sentences / total number of sentences. Table 7 shows results for 12 linguistic structure types in human and machine translations.

Table 7: Data on 12 linguistic structure types in human and machine translations

	DeepL	Baidu	Youdao	Tencent (name)	Human
word length	3.54125	3.45784	3.46039	3.83025	3.17815
length of a sentence	25.6494	28.2785	28.1351	24.1811	30.8995
type-token ratio	303.000	298.500	300.833	278.833	338.667
proportion of adverbs	0.0137514	0.0150754	0.0182825	0.0155409	0.0265748
percentage of nouns	0.315732	0.319375	0.324654	0.346085	0.302657
proportion of verbs	0.217822	0.219989	0.208864	0.237298	0.207677
proportion of pronouns	0.00165017	0.00223339	0.000554017	0.00119546	0.00295276
proportion of conjunctions	0.00220022	0	0.000554017	0.00179319	0.000984252
proportion of adjectives	0.0561056	0.0625349	0.0581717	0.0645547	0.0497047
proportion of declarative sentences	0.880478	0.876712	0.995495	0.886792	0.870813
exclamatory sentence ratio	0.0159363	0.0273973	0.0045045	0.0150943	0.0287081
proportion of questions	0.103586	0.0958904	0	0.0981132	0.100478

Although word length in Chinese is determined by the number of characters and in English by the number of letters, both languages share a commonality: the longer the word, the less semantic content it tends to carry [9], resulting in a more formal and less flexible style. Table 7 shows Tencent has the largest word length, followed by DeepL, Youdao and Baidu; all machine translations have longer word lengths than human translation, indicating less flexible expression. In addition, compared with other machine translations, Tencent's expression flexibility needs to be improved, while Baidu's and Youdao's expression flexibility is closer.

For sentence length, longer sentences typically mean more formal style, greater structural complexity, and higher language proficiency. Human translation has the longest sentences, while Tencent's are the shortest. Human translations significantly surpass machine translations in sentence length, indicating higher sentence complexity and language proficiency. Baidu and Youdao have the most similar sentence lengths, suggesting similar levels of language proficiency and sentence complexity. DeepL, Baidu, and Youdao all exceed Tencent by 4 percentage points, highlighting the need for Tencent to improve its language proficiency and sentence complexity.

The type-token ratio (ratio of word count to lexical properties) reflects vocabulary diversity: a higher value indicates richer, more varied vocabulary in the translation. The type-token ratio of human translations is the highest, followed by DeepL, Youdao and Baidu, and Tencent is the lowest. Human translation has a type-token ratio of 338.667, compared to Tencent's 278.833 with a 59.834 difference (17.67%), indicating significantly larger vocabulary size, greater richness and variability in human translation, and higher word repetition in machine translation.

In word classes, human translations have significantly higher adverb proportions than individual

machine translations, indicating that the linguistic richness of the human translations is much higher than that of the machine translations; the proportions of nouns, verbs and adjectives in the machine translations are higher than those of the human translations but the difference is not significant, indicating that the machine translations are close to the human translations in terms of the ability of translating nouns, verbs and adjectives; and the proportions of pronouns and conjunctions in the human translations are all higher than those of the individual machine translations, indicating that the human translations are better than machine translations in terms of pronoun reduction and translation logic.

In terms of sentence types, machine translations contain a higher density of declarative sentences than human translations, with insignificant differences. This is likely because their slightly lower proportions of other sentence types increase the share of declarative ones, indicating a similar ability to handle such sentences. In terms of exclamatory sentences, all the human translations are significantly higher than those of the machine translations. In interrogative sentences, human translations exceed other machine translations, while DeepL is slightly higher but with insignificant difference.

To further compare how machine translations and human translations grasp the sentiment of the source language, this study presents a sentiment analysis frequency chart for major machine translations and human translations, see Figure 2.

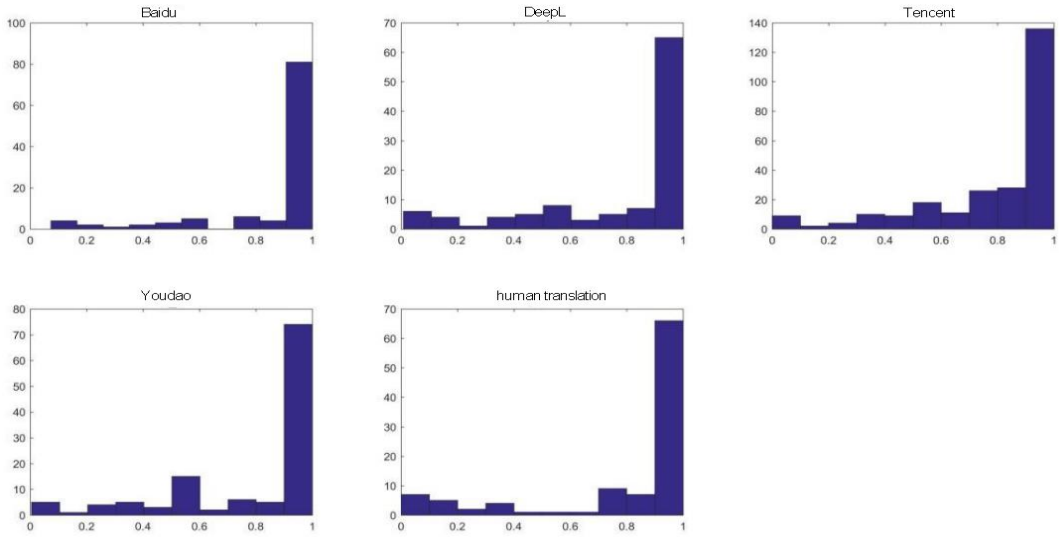

Figure 2: Frequency plot of statement sentiment analysis

In Figure 2, the horizontal axis represents utterance sentiment intensity (values closer to 0 indicate stronger negativity, closer to 1 stronger positivity), and the vertical axis denotes the number of sentences corresponding to each intensity level. As shown in Figure 2, human translations cluster near 0 within the 0.4-0.6 range, whereas Baidu, DeepL, Tencent, and Youdao exhibit more pronounced values in this interval. This suggests that machine translations capture sentiment with less clarity than human translations, while human translations display greater subjectivity in their sentiment analysis.

6. Conclusion

This study compares machine and human translations using an LSTM neural network model and deep neural network translation software. The findings reveal that human translation outperforms machine translation significantly across fields, with a narrower gap in daily language. This suggests that daily communication, with its simplicity and lack of complex sentence structures, enables

machine translation to achieve high fidelity and clarity. In literary contexts, however, human translation significantly outperforms machine translation, indicating that translation demands not just fidelity and clarity but also elegance, which is the key factor behind the substantial accuracy gap between machine and human translations in literary works. In fields like economy, trade, and business, which are marked by extensive specialized terminology, machine translation's failure to promptly update its text databases may reduce translation accuracy.

Furthermore, this study analyzes 12 linguistic structure types, including word length, sentence length, and type-token ratio, to compare the translation performance of human and machine translations in the contexts of business negotiations, literature, and speech addresses. The findings indicate that human translations excel over machine translations in terms of flexibility of expression, sentence complexity, language proficiency, translation logic, vocabulary richness, and variability. Additionally, this research conducts a comparative analysis of sentiment analysis between machine and human translations, revealing that human translations are better at recognizing the tone and emotions of the source text. In professional domains and literary works, though machine translations in this study's samples have relatively lower accuracy, they still reach nearly half that of human translations, effectively assisting translators and significantly reducing their workload.

In sum, while machine translation cannot fully replace human translation at present, it can assist translators in various fields and reduce their workload. Future enhancements should prioritize deep semantic comprehension, flexible expression, and sentiment analysis to strengthen human -machine collaboration.

References

[1] Dai G R & Liu S Q. Neural network machine translation: progress and challenges[J]. Foreign Language Teaching, 2023(1):82.

[2] Hou Q & Hou R L. Neural machine translation research-insights and prospects[J]. Journal of Foreign Languages, 2021(5):54.

[3] Fan W Q & Wang Y. Spiritual interaction between translator and text: the bottleneck of machine translation[J]. Theory and Practice of Foreign Language Teaching, 2022(3):137.

[4] Yan C S. Machine translation replacing human translation is still just a vision--review of examples of machine translation from English to Chinese[J]. Translation Teaching and Research, 2022(1):12.

[5] Poibeau T. Machine Translation[M]. Boston: The MIT Press, 2017.

[6] Huang W & Liu H T. Application of measurement features of Chinese corpora in text clustering[J]. Computer Engineering and Applications, 2009(29):26.

[7] Chen X Y, Li W W & Wang Y. The application of measurement characteristics in comparison of language styles and determination of writers--taking Han han's "Threefold Gate" and Guo Jingming's "In Dreams" as an example[J]. Computer Engineering and Application, 2012(3):137.

[8] Jiang Y. Comparison of linguistic measurement characteristics between manual translations and machine online translations--taking the 5 sessions of Han Suyin translation contest English-Chinese manual translations and online translations as an example[J]. Foreign Language Teaching, 2014(5):100.

[9] Zipf G K. The psycho-biology of language: an introduction to dynamic philology[M]. London: G. Routledge &Sons Ltd., 1936.