Taylor & Francis
Taylor & Francis Group

Applied Artificial Intelligence
An International Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

An Analysis of the Evaluation of the Translation Quality of Neural Machine Translation Application Systems

Shanshan Liu & Wenxiao Zhu

To cite this article: Shanshan Liu & Wenxiao Zhu (2023) An Analysis of the Evaluation of the Translation Quality of Neural Machine Translation Application Systems, Applied Artificial Intelligence, 37:1, 2214460, DOI: 10.1080/08839514.2023.2214460

To link to this article: https://doi.org/10.1080/08839514.2023.2214460

Published online: 23 May 2023.

APPLIED ARTIFICIAL INTELLIGENCE
2023, VOL. 37, NO. 1, e2214460 (1531 pages)
https://doi.org/10.1080/08839514.2023.2214460

Taylor & Francis
Taylor & Francis Group

An Analysis of the Evaluation of the Translation Quality of Neural Machine Translation Application Systems

Shanshan Liu and Wenxiao Zhu

College of Foreign languages, Henan University of Chinese Medicine, Zhengzhou, China

ABSTRACT

Neural machine translation (NMT) is applied to generate a more reliable and accurate translation practice as the most cutting-edge technology. In recent years, NMT has achieved gratifying results. However, the main obstacle for market-oriented NMT application systems appears to suffer from weak translation quality that fails to meet users' needs. This paper focuses on the machine translation of political documents and implements six dominant NMT application systems in the market to evaluate their translation quality. The evaluation process further employs both BLEU and NIST technical evaluation algorithms and re-verifies the results with the manual evaluation method called the "Score Ranking System" to compare the performances of the six NMTs in Chinese-English translations of political documents. Through diagnosis and evaluation of the problems and errors in NMTs, the paper eventually proposes the "Cue Lexicon+" model to remedy prominent problems. Besides, the "NMT+ Lexicon Intelligent Translation Assistant" soft is developed and the "Cue Lexicon+" is integrated into the NMT application systems to further improve the translation quality, providing a reference and research basis to increase the performance and update the NMT application systems.

ARTICLE HISTORY
Received 20 March 2023
Revised 28 April 2023
Accepted 29 April 2023

Introduction

Neural network technologies help machine translation to increasingly mature. NMT application systems are widely used and have yielded huge benefits. Several major technology companies have started to develop their own NMT practical systems (Zhou et al. 2016). In 2013, Baidu (Sun and Kumar 2022) began to research NMT; in 2015, it took the lead in adopting the deep neural network in the machine translation system. Afterward, the machine translation quality was significantly improved (Sun and Kumar 2022). In 2017, Google proposed the Transformer model. Many excellent pre-trained language models and machine translation models were developed, such as the BERT and the GPT series, which constantly refreshed the ability level of many natural language processing tasks (Wu et al. 2016). In 2019, Volctrans

CONTACT Wenxiao Zhu zhuwx1128@163.com College of Foreign languages, Henan University of Chinese Medicine, Zhengzhou 450046, China

proposed LightSeq. The first open-source engine could fully support the high-speed inference of various models such as Transformers and GPT in the industry. In LightSeq, the Transformer-based sequence feature extractor (Encoder) and autoregressive sequence decoder (Decoder) were further optimized.

In the 2020 Conference on Machine Translation (WMT20), Volctrans won the championship in the "Chinese-English" language translation contest with a significant advantage among 39 participating teams (Wu et al. 2020). Also, the DeepL which is known as the most accurate translation artifact in the world was presented on its official website (Yulianto and Supriatnaningsih 2021). The continuous improvement of the performance of NMT application systems is inseparable from the research on the evaluation of machine translation quality. Ultimately, the evaluation of machine translation quality is a linguistic issue of comparing sentences; therefore, scholars must combine machine translation with linguistic research (Guzmán et al. 2017). At present, most scholars focus on the evaluation of English Chinese machine translation. However, there are still few papers on the overall quality performance. Especially, Chinese-English t of political documents are rare.

Therefore, the manuscript utilizes six dominant NMT application systems (Google, DeepL, Amazon, Baidu, Volctrans, and IFLYTEK), and carries out comparative research on machine translations and manual translations by combining quantitative evaluation with qualitative evaluation. It analyzes the problems and errors in machine translations and puts forward targeted improvement schemes based on technical evaluation and manual evaluation. Moreover, this paper will shed light on the choice of the relatively best NMT application system for political documents and the improvement steps in the performance of NMT application systems.

The Design of the Research

Selection of Research Objects

Aiming to investigate the overall quality performance of NMT application systems in the Chinese-English translation of political documents, the paper chose the first and second volumes of "Xi Jinping: The Governance of China" as source language texts (STs) (see Table 1) according to the following principles: (1) Reality: to choose only authentic natural STs to investigate the natural text processing abilities of major NMT application systems; (2) Moderate difficulty: to exclude text materials that were too simple or difficult because they could not correctly reflect the actual levels of the NMT application systems; (3) Single style: to research only political documents represented by "Xi Jinping: The

Table 1. List of texts in the source language.

Case No.	Title	Number of Chinese Characters	Number of Segments
Case 1	"Xi Jinping: The Governance of China I"	188,213	5,244
Case 2	"Xi Jinping: The Governance of China II"	241,962	5,079
Total		**430,175**	**10,323**

Governance of China" to ensure their representativeness; (4) Official standard translations: to evaluate the translation quality of the six NMT application systems by referring to the comparative translations.

Research Evaluation Methods

The translation quality evaluation is an essential step in improving the performance of translation systems. The qualitative and quantitative evaluations of translation quality are two aspects to assess translation quality. The former is the basis and principle of the latter, while the latter is the objective and digital result of the former. To rephrase, they are inseparable (Duh 2008). In this section, quantitative evaluation and qualitative evaluation will be combined to ensure that the score obtained can truly reflect the translation quality level, and to provide a translation quality evaluation scheme for NMT application systems.

Technical Evaluations of BLEU and NIST

In the machine translation field, technical evaluation is the usual method. The evaluation system compares the machine-translated text automatically with the reference translation. A final score is generated. The dominant evaluation methods are called BLEU and NIST.

BLEU evaluation index: BLEU (Bilingual Evaluation Understudy) is an evaluation index for evaluating machine translation results, and its value ranges from 0 to 1. The closer it is to 1, the closer the machine translation result is to the reference translation; the closer it is to 0, the more the machine translation result deviates from the reference translation (Mathur, Baldwin, and Cohn 2020). BLEU uses accuracy to measure the length of the machine translation result approaching reference translation. When calculating the accuracy, the number of n consecutive sequence matches between the machine translation results and the reference translation must be first known. More matches indicate a higher BLEU value, which means that the machine translation result is more like the reference translation. Eq. (1) presents the number of n consecutive matches,

$$Count_{clip}(n-gram) = \min(Count, Max\_ef\_ount)$$

where $n-gram$ denote n consecutive sequences; $Count$ represents the total number of $n-gram$ occurrences in the machine translation result;

Max_ef_ount denotes the total number of $n-gram$ occurrences in the reference translation; $Count_{clip}(n-gram)$ denotes the number of matches between the numbers of occurrences of n consecutive sequences in the machine translation and the reference translation.

After obtaining the number of consecutive matches, the translation accuracy can be calculated. The Eq. (2) presents the precision,

$$precision_n = \frac{\sum_{n-grameC} Count_{clip}(n-gram)}{\sum_{n-grameC} Count(n-gram)}$$

where $Count_{clip}(n-gram)$ denotes the number of occurrences between the numbers of occurrences of n consecutive sequences in the machine translation result and the reference translation; $Count(n-gram)$ represents the total number of occurrences of n consecutive sequences in the machine translation.

Since the length of the machine translation is less than the length of the reference translation, the BLEU score will be affected. In this case, a penalty factor will be introduced to control the issue. Therefore, a length penalty factor (Brevity penalty factor) is introduced. Eq. (3) presents the BP,

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

where $r$denotes the total length of the reference translation; c denotes the total length of the machine translation result; BP represents the penalty factor.

The BLEU evaluation index is shown in Eq. (4) as follows:

$$BLEU = BP \times \exp\left(\sum_{n=1}^{N} w_n \log \text{precision}\, n_n\right)$$

where N represents the maximum order of n continuous sequences; w_n represents the weight coefficient.

Since the overall translation accuracy gradually decreases with the increase of N, N generally is set to 4. However, the BLEU has its setbacks. To be specific, it focuses on the details of the sentences but neglects the coherence of the overall translation.

NIST evaluation index: NIST is an improvement based on the principles of BLEU. It adds weights to different words in sentences to emphasize the translation of key semantics. The index adds more weights to word sequences containing more information. Eq. (5) presents the computation of weights used in the NIST evaluation index,

$$weight(w_1 \ldots w_n) = \log \frac{count(w_1 \ldots w_{n-1})}{count(w_1 \ldots w_n)}$$

where $count(w_1 \ldots w_{n-1})$ denotes the number of occurrences of n-1 consecutive word sequences in the reference translation; $count(w_1 \ldots w_n)$ represents the number of occurrences of n consecutive word sequences in the reference translation; $weight(w_1 \ldots w_n)$ denotes the final weight of n consecutive word sequences.

The NIST evaluation index can be modified further based on the BLEU evaluation index, which is shown in Eq. (6),

$$precision\_nist(n) = \frac{\sum_{n-grame \in C} weight(n-gram) \cdot Count_{clip}(n-gram)}{\sum_{n-gram \in C} Count(n-gram)}$$

where $Count_{clip}(n-gram)$ denotes the number of occurrences between the numbers of occurrences of n consecutive word sequences in the machine translation result and the reference translation; $Count(n-gram)$ represents the total number of occurrences of n consecutive word sequences in the machine translation result; $weight(n-gram)$ denotes the final weight of the n consecutive word sequences.

The final NIST evaluation index is shown in Eq. (7) as follows:

$$NIST = BP \times \exp\left(\sum_{n=1}^{N} \log precision\_nist(n)\right)$$

where BPdenotes the penalty factor; $precision\_nist(n)$ represents the translation accuracy.

Evaluation of Manual Score Ranking

The high-reliability manual scoring is the key to building an automatic scoring system for Chinese-English translation. The paper follows the Chinese-English translation principles of "Fidelity" and "Fluency" (Feng et al. 2020) and adopts the manual scoring criteria in China's 863 program in machine translation evaluation mentioned by Reiss (Reiss and Rhodes 2014). The criteria offered a framework for evaluating the machine translation quality of this paper. Both fidelity and fluency are the primary criteria of the evaluation system throughout the research, which has six evaluation levels and a scale changing from 0 to 5 (corresponding to scores 0 through 5). The evaluation results were recorded up to the second digit after the decimal point to ensure objectivity. The scoring criteria for fidelity and fluency are shown in Table 2:

The machine translation quality is always evaluated with a real number score. However, different understandings of the evaluation criteria and scales for manual translation quality may result in poor consistency and instability of the evaluation results (Ghorbani et al. 2021). The ranking method was more

Table 2. The fidelity and fluency scoring criteria for manual evaluation of machine translations.

Level	Scoring Criteria	Scoring Criteria
0	The text is not translated at all	The translation is completely incomprehensible
1	Only a few words in the translation match ST	The translation is obscure and difficult to understand
2	Only a few contents in the translation match ST	The translation is not fluent
3	The translation matches ST	The translation is fluent
4	The translation expresses most of the information in the ST	The translation is fluent, but it lacks features of the native language
5	The translation accurately and completely expresses the information in ST	The translation is fluent and authentic

reliable than the scoring method for evaluating the quality of multiple machine translations was generally proposed (Shapran et al. 2021).

The research underwent meticulous designs in the scoring mechanism and operations to obtain high-reliability scoring results and ensure the quality of comparative research on translations. The "Score Ranking System" evaluation method was designed to ensure effective statistics and analysis on all scores, further, ranking was performed by researchers based on the scores to avoid the distortion of results due to the differences in raters' scores. The raters consisted of nine graduates with a master's degree in translation and possessed a Level II certificate from China Accreditation Test for Translators and Interpreters (CATTI). The raters were divided into three groups with three persons in each group. The 300 segments sampled were equally divided into three groups, each of which contained 100 segments. The 100 segments of each group were scored and ranked by the raters in each group. The six NMT application systems were ranked from 1st to 6th based on their translation quality. They were presented in Tables 3 and 4, which are ranking statistics. So, more objective, and fair results were achieved using the ranking method (the implementation steps of Case 2 are the same as above). Afterward, the descriptive

Table 3. Ranking statistics.

	Google	Deepl	Amazon	Baidu	Volctrans	IFLYTEK
Ranking results by Rater 1 in Group A						
Segment 1	2	1	2	3	2	2
Segment 2	3	1	5	4	6	2
Segment 3	2	2	3	1	1	1
……	……	……	……	……	……	……
Ranking results by Rater 2 in Group A						
Segment 1	1	3	2	2	4	1
Segment 2	1	3	6	5	4	2
Segment 3	2	3	4	1	1	1
……	……	……	……	……	……	……
Ranking results by Rater 3 in Group A						
Segment 1	2	4	3	3	3	1
Segment 2	2	3	4	1	1	1
Segment 3	3	3	4	1	2	1
……	……	……	……	……	……	……

Table 4. Statistics of the totalized data of rankings made by raters by segment.

	Google	Deepl	Amazon	Baidu	Volctrans	IFLYTEK
Summation of the values of rankings made by the three raters in Group A by segment						
Segment 1	5	8	7	8	9	4
Segment 2	6	7	15	10	11	5
Segment 3	7	8	11	3	4	3
……	……	……	……	……	……	……

statistics of the scores and rankings were carried out with SPSS 23 version. The raters were required to strictly follow the scoring rules and to take notes of the translation segments that appeared problematic to ensure objectivity and consistency in the formal scoring process. The implementation process of the "Score Ranking System" evaluation method is presented in Figure 1.

The Steps of Specific Research

Research steps (see Figure 2): STs and their official reference translations (RTTs) were converted into the Chinese-English text segments concurrently (see Figure 3). The STs were imported into the NMT application systems of Google, DeepL, Amazon, Baidu, Volctrans, IFLYTEK to generate target language texts (TTs). STs and the corresponding TTs were copied into WORD documents. Numbers were given, and documents were archived (see Figure 4). The technical evaluation, which was subject to the quantitative analysis, mainly adopted BLEU and NIST to evaluate the TT quality of the six NMT application systems. The manual evaluation, which was subject to the qualitative analysis, verified the technical evaluation results for the second time and adopted the "Score Ranking System" to evaluate the TT quality of the six NMT application systems.

Evaluation and Result Analysis

Case 1: Quality Evaluation Results of Machine Translations of Xi Jinping: The Governance of China I

Technical Evaluation Scores

The Result of the BLEU Evaluation. The BLEU scores of the six NMTs appeared from high to low, which is IFYTEK>Baidu>Google>Volctrans>DeepL>Amazon (see Figure 5). The TT scores of IFYTEK appeared 0.5391 in 1-gram, 0.3868 in 2-gram, and 0.2798 in 3-gram. All are higher than the other five NMT application systems. Namely, the machine translation of IFYTEK had the highest matching degree of N-grams with the reference translation. All words in the translation had the greatest contribution to the meanings, and the translation was more fluent and readable, while the Amazon translation was the opposite.

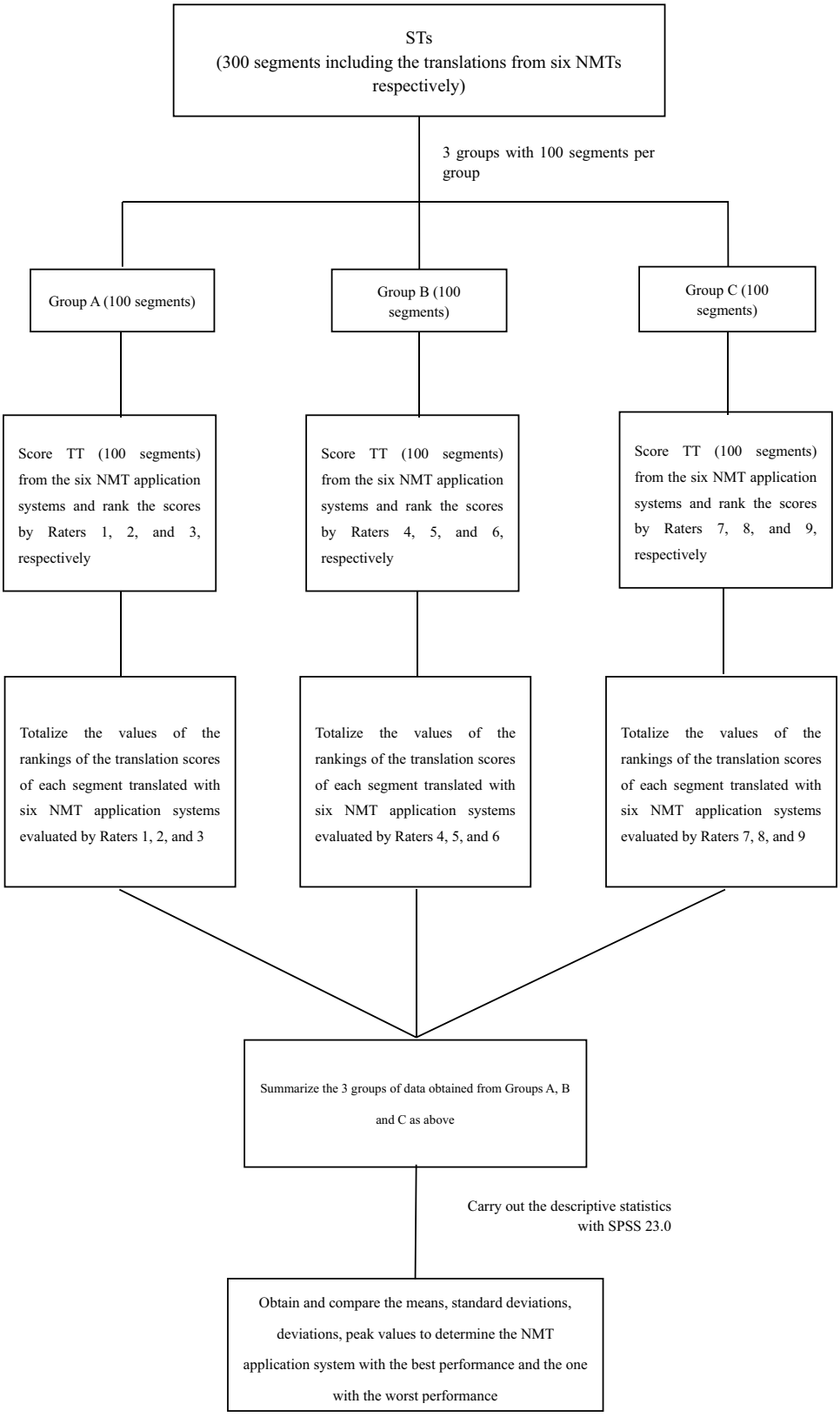

Figure 1. Flow chart of evaluation method called "Score Ranking System.

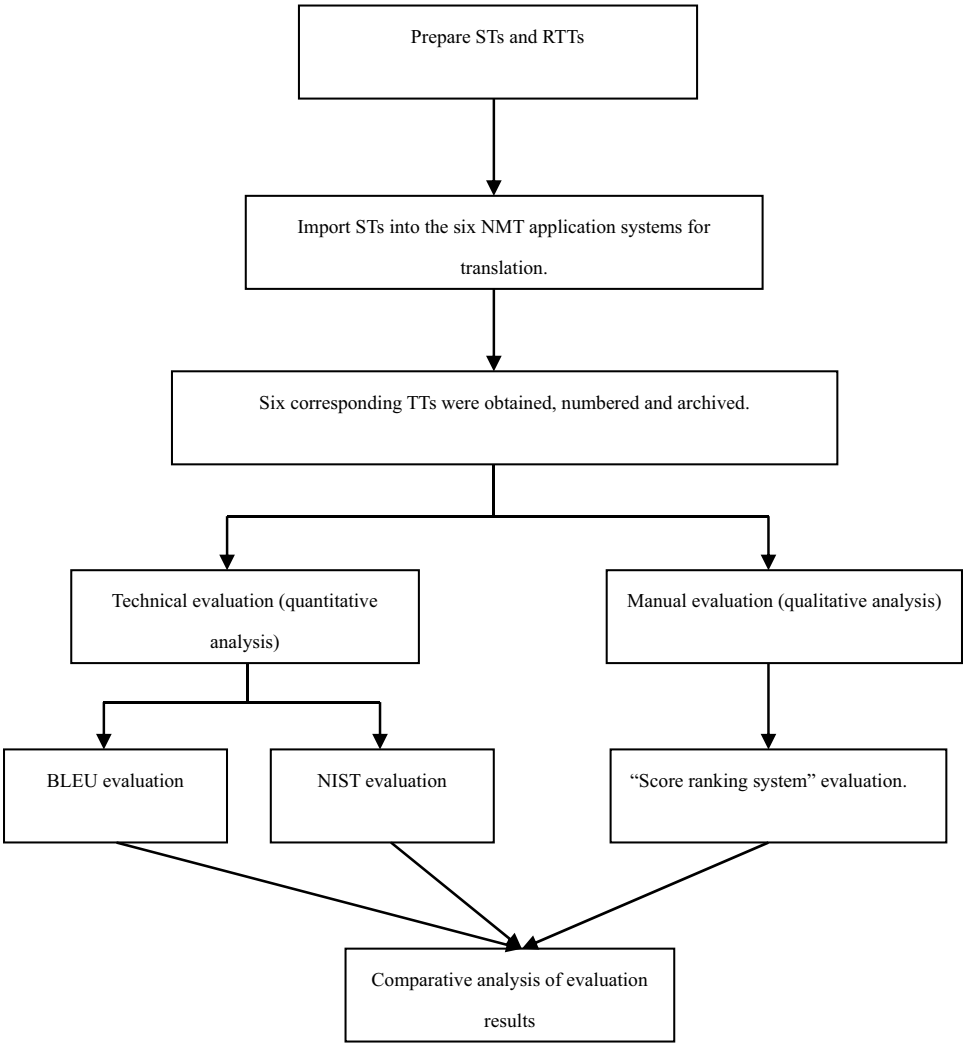

Figure 2. The flow chart of the research steps.

The Result of the NIST Evaluation. The order of NIST scores from high to low is presented as follows: IFYTEK>Baidu>Google>Volctrans>DeepL>Amazon (see Figure 6). Since NIST is the total amount of the information obtained divided by the number of n-gram segments in the entire translation, the weight of the keywords with low occurrence frequency could be increased, that is, the key words had low occurrence frequency. The 1-gram score was 4.3414, the 2-gram was 5.6736, the 3-gram was 6.0477, and the n-gram had higher accuracy in the IFYTEK translation system. So, the scores for correct consecutive translations were also higher than those in the other five NMT application systems, that is, IFYTEK's TTs contained more information for each sentence, and had the highest overall translation quality. The results obtained from the two machine translation evaluation technologies were

Segment-by-segment contrast version between ST and RTT		
No.	**STs**	**Official TTs**
Segment 1	习近平谈治国理政第一卷	Xi Jinping: The Governance of China I
Segment 2	一、坚持和发展中国特色社会主义	I. Socialism with Chinese Characteristics
Segment 3	人民对美好生活的向往，就是我们的奋斗目标	The People's Wish for a Good Life Is Our Goal
Segment 4	这是习近平在十八届中央政治局常委同中外记者见面时讲话的主要部分。	Part of the speech at the press conference by members of the Standing Committee of the Political Bureau of the 18th CPC Central Committee.
Segment 5	记者朋友们对十八大作了大量报道，向世界传递了许多"中国声音"。	Friends from the news media have extensively covered the 18th National Congress of the Communist Party of China (CPC), conveying to the world many voices from China.
Segment 6	我代表大会秘书处向大家表示衷心的感谢。	On behalf of the Secretariat of the Congress, I wish to express our sincere thanks to you.
Segment 7	刚才，我们召开了中国共产党第十八届中央委员会第一次全体会议，选举产生了新一届中央领导机构，选举我为中央委员会总书记。	We have just held the First Plenary Session of the 18th CPC Central Committee, and elected a new central leadership. I was elected general secretary of the Central Committee.
Segment 8	我代表新一届中央领导机构成员感谢全党同志的信任，定当不负重托，不辱使命。	On behalf of the members of the newly-elected leadership, I wish to express our thanks to all other members of the Party for their trust in us. We will do our utmost to be trustworthy and fulfill our mission.
……	……	……
……	……	……

Figure 3. Segment-by-segment contrast version between ST and RTT.

TTs of six NMT application systems							
No.	**STs**	**Google TTs**	**Deepl TTs**	**Amazon TTs**	**Baidu TTs**	**Volctrans TTs**	**IFYTEK TTs**
Segment 1	习近平谈治国理政第一卷	Xi Jinping talks about the governance of the country, volume 1	Xi Jinping on Governance, Volume 1	Xi Jinping Talks About Governance and Governance Volume 1	Xi Jinping on Governance Volume I	Xi Jinping on State Governance Volume 1	Xi Jinping: The Governance of China Volume I
Segment 2	一、坚持和发展中国特色社会主义	1. Uphold and develop socialism with Chinese characteristics	I. Adherence to and development of socialism with Chinese characteristics	I. Adhere to and develop socialism with Chinese characteristics	1、 Adhere to and develop socialism with Chinese characteristics	1. Uphold and develop socialism with Chinese characteristics	I. Adhering to and Developing Socialism with Chinese Characteristics
Segment 3	人民对美好生活的向往，就是我们的奋斗目标	People's yearning for a better life is our goal	The people's desire for a better life is our goal	The people's yearning for a better life is our goal	To meet their desire for a happy life is our mission	The people's yearning for a better life is our goal	People's yearning for a better life is our goal of struggle.
Segment 4	这是习近平在十八届中央政治局常委同中外记者见面时讲话的主要部分。	This is the main part of Xi Jinping's speech when he met with Chinese and foreign journalists at the Standing Committee of the Political Bureau of the 18th CPC Central Committee.	This is the main part of Xi Jinping's speech at the meeting of the Standing Committee of the 18th Central Political Bureau with Chinese and foreign journalists.	This is the main part of Xi Jinping's speech during a meeting with Chinese and foreign reporters at the 18th Politburo Standing Committee of the CPC Central Committee.	This is the main part of Xi Jinping's speech when the Standing Committee of the Political Bureau of the 18th Central Committee met with Chinese and foreign journalists.	This is the main part of Xi Jinping's speech when the Standing Committee of the Political Bureau of the 18th Central Committee met with Chinese and foreign journalists.	This is the main part of Xi Jinping's speech when he met with Chinese and foreign journalists at the 18th Standing Committee of the Political Bureau of the Central Committee.
……	……	……	……	……	……	……	……
……	……	……	……	……	……	……	……

Figure 4. Tts of six NMT application systems.

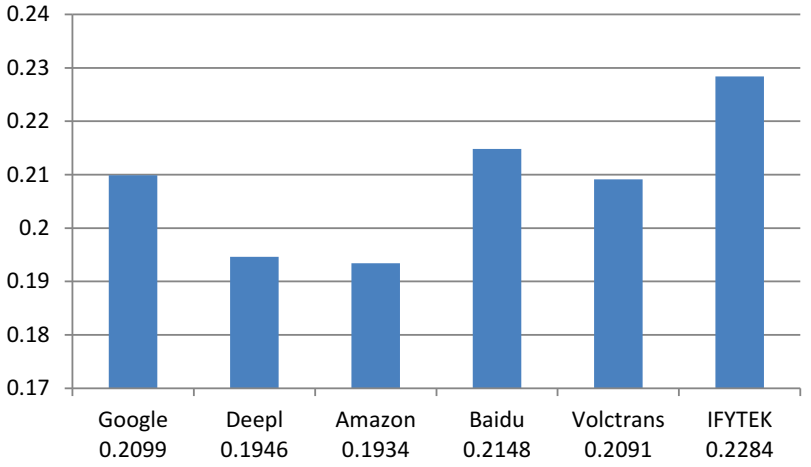

Figure 5. BLEU evaluation scores of TTs translated with six NMTs.

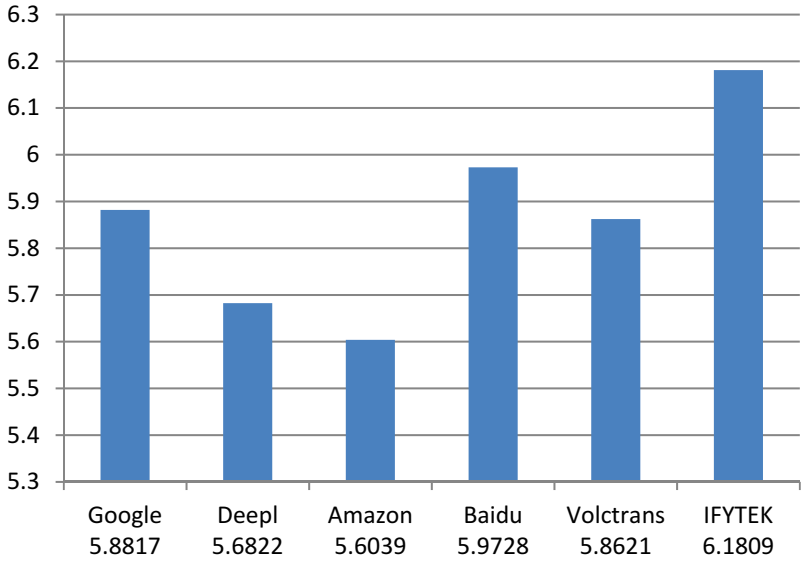

Figure 6. NIST evaluation scores of TTs translated with six NMTs.

found to be consistent. While the highest score was given to IFYTEK, the lowest one was given to Amazon.

The Result of Manual Evaluation

After ranking the scoring results of each segment in the 300 segments sampled from the TTs generated by the six NMT application systems with the "Score Ranking System," the overall descriptive statistics were calculated with SPSS 23 version. Table 5 presents them.

Table 5. Descriptive statistics of scores of six NMTs.

	Number of cases	Min	Max	Sum	Mean	Standard Deviation
Google	300	3.0000	16.0000	2183.0000	7.276667	2.7468661
DeepL	300	3.0000	17.0000	2351.0000	7.836667	2.8875217
Amazon	300	3.0000	16.0000	2405.0000	8.016667	2.9876651
Baidu	300	3.0000	15.0000	2070.0000	6.900000	2.5998456
Volctrans	300	3.0000	16.0000	2238.0000	7.460000	2.7834436
IFYTEK	300	3.0000	16.0000	2042.0000	6.806667	2.6291310
Number of valid cases	300					

According to the comparison of the sums and the means of the TT scores of the six NMT application systems, the order was found to be IFYTEK＜Baidu ＜Google＜Volctrans＜DeepL＜Amazon. Thus, the smaller the value, the higher the ranking, that is, IFLYTEK ranked first, with a mean of about 6.81 and a standard deviation of 2.63; Amazon ranked last, with a mean of about 8.02 and a standard deviation of 2.99. It was seen that the ranking values of Amazon were more polarized, so its overall performance was more unstable leading to the relatively worst translation quality.

In conclusion, the results of BLEU, NIST, and manual evaluations of TTs translated with the six NMT application systems were completely consistent. IFLYTEK had the best evaluation result and overall translation quality.

Case 2: Quality Evaluation Results of Machine Translations of Xi Jinping: The Governance of China II

Technical Evaluation Scores

The Result of the BLEU Evaluation. The BLEU scores were presented from high to low as follows: IFYTEK>Baidu>Google>Volctrans>DeepL>Amazon (see Figure 7). The TT scores of IFYTEK appeared 0.5387 in 1-gram, 0.3833 in 2-gram, and 0.2879 in 3-gram. All are higher than those of the five NMT application systems. Namely, the TTs and RTTs of IFYTEK had the highest matching degree of N-grams, the words, the syntax, and segments were the most similar, while Amazon translations were found to be the opposite.

The Result of the NIST Evaluation. The NIST scores were presented from high to low as follows: IFYTEK>Baidu>Google>Volctrans>DeepL>Amazon (see Figure 8). It was seen that the TT scores of IFYTEK appeared 4.3108 in 1-gram, 5.5873 in 2-gram, and 5.9159 in 3-gram, and n-gram had higher accuracy in IFYTEK translation system, with higher scores for correct consecutive translations than those of the five NMT application systems, that is, IFYTEK’s TTs contained more information for each sentence, and had the highest overall translation quality. The results obtained from the two machine translation evaluation technologies were consistent, with the highest score being given to IFYTEK and the lowest score to Amazon.

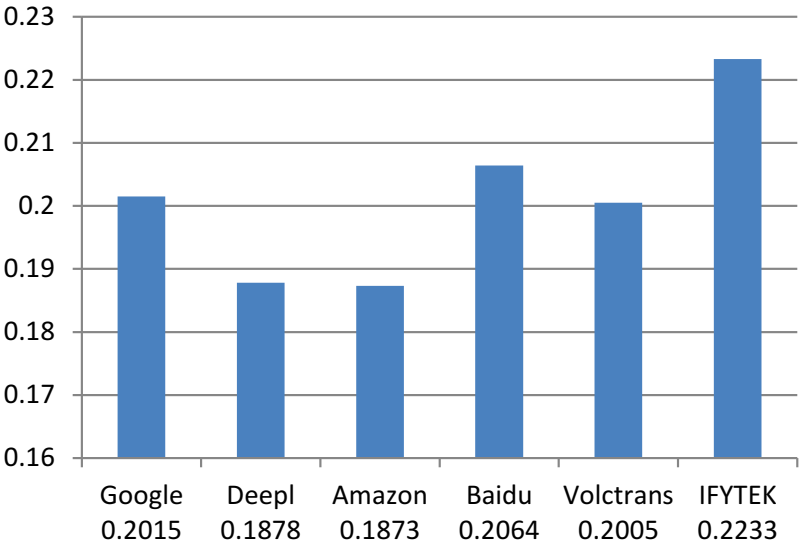

Figure 7. BLEU evaluation scores of TTs translated with six NMTs.

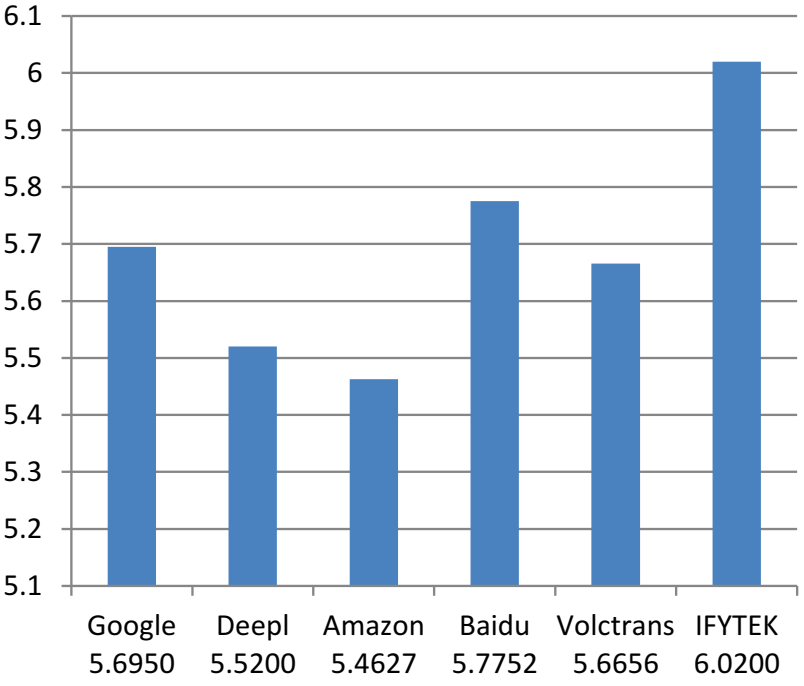

Figure 8. NIST evaluation scores of TTs translated with six NMT application systems.

Manual Evaluation Scores

After ranking the scoring results of each segment in the 300 segments sampled from the TTs generated by the six NMT application systems with the "Score

Table 6. Descriptive statistics of scores of six NMTs.

	Number of valid cases	Min	Max	Sum	Mean	Standard Deviation
Google	300	3.0000	15.0000	2534.0000	8.446667	2.4550451
DeepL	300	3.0000	16.0000	2668.0000	8.893333	2.5853890
Amazon	300	3.0000	17.0000	2724.0000	9.080000	2.7723431
Baidu	300	3.0000	15.0000	2460.0000	8.200000	2.4806947
Volctrans	300	3.0000	15.0000	2628.0000	8.760000	2.3467903
IFYTEK	300	3.0000	14.0000	2379.0000	7.930000	2.3799117
Number of valid cases	300					

Ranking System," the overall descriptive statistics are computed with SPSS 23 version. Table 6 depicts the outcomes.

According to the comparison of the means of the TT scores of the six NMT application systems, the order was found to be IFYTEK＜Baidu＜Google＜Volctrans＜DeepL＜Amazon. Namely, IFLYTEK ranked first, with a mean of about 7.93 and a standard deviation of 2.38; Amazon ranked last, with a mean of about 9.08 and a standard deviation of 2.77. It was seen that the ranking values of Amazon were more polarized, so its overall performance was more unstable, with the worst translation quality.

In conclusion, the results of BLEU, NIST, and manual evaluations of TTs translated with the six NMT application systems were completely consistent. IFLYTEK had the best evaluation result and overall translation quality.

Summary of Quality Evaluation

The results of technical evaluation and manual evaluation on Cases 1 and 2 are completely consistent and mutually verified. Concluded that IFLYTEK has the best performance in the evaluations. Therefore, IFYTEK possesses the best translation quality among the six NMT application systems in terms of Chinese-English translations of political documents.

Discussions

To further analyze the problems with machine translation, a manual ranking of scores was performed covering all segments (10,323 segments) of Cases 1 and 2 that were categorized and annotated based on types of problems by counting and inferring the percentages of each type of error. The types of machine translation errors identified in the comparison process were classified regarding the secondary classification system. The generalization of the error categories framework was adopted using the research by Lu and Li (Luo and Li 2012). The error types involved were added or deleted by combining them with the actual case text comparison. The errors in machine translations were classified into three types first-order errors based on lexicon, syntax, and so on. The first-order errors were further classified into second-order ones

Table 7. Types of errors in machine translations.

First-order errors	Second-order errors
1. Lexical errors	1.1 Errors in Terminology
	1.2 Wrong part-of-speech or tense
	1.3 Improper use of words or incorrect collocation
2. Syntactic errors	2.1 Wrong segmentation of a long difficult sentence
	2.2 Mistranslation of non-subject sentences
	2.3 Confusion in sentence structure
3. Other errors	3.1 Improper Cohesion and Coherence
	3.2 Errors in capital and small letters
	3.3 Omissions

according to the characteristics of specific errors shown in Table 7. The error rates of different machine translations in Table 8 were counted after a segment-by-segment comparison between the translations of NMT application systems and human translations, as well as error annotation. The translation error refers to at least one type of error in machine translations listed in Table 7. The error rate is one-tenth of the result of the total number of segments of faulty machine translations divided by the total number of segments of machine translations. The errors of the same type, for instance, improper use of words that occurred repeatedly in the same segment were annotated and counted only once, and errors of different types in the same segment were annotated and counted separately. Errors in machine translations found in the comparative analysis will be described later.

Lexical Error

Words are the basic elements that constitute a sentence, and vocabulary translation has a crucial influence on the quality of translations (Catford 1978). Table 8 depicts that the rate of lexical errors is much higher than that of other types of errors, up to 69.28% as many sentences are composed of the most basic element of vocabulary. In the type of high-frequency lexical errors, the error rate of terminology tops the list, accounting for 63.14% of 600 machine-translated sentences. This indicates the major linguistic feature of using many terminologies in political documents. The error rate for the wrong part of speech or tense is 4.41%, while that of improper use of words or collocation is 1.7%, accounting for the minimum percentage.

(1) Terminology error. ST: “照此说来，博鳌亚洲论坛正处在一个新的起点上，希望能更上一层楼。” In this sentence, “博鳌亚洲论坛” was translated differently in different databases. It was translated into “the Bodie Forum for Asia,” “the BoF,” “the Bo Turtle Asia Forum,” “the Bo Bie Asia Forum,” “the Bobian Asia Forum,” and “the Boyi Asia Forum” by Google, DeepL, Volctrans, IFLYTEK, Baidu, and Amazon respectively. The translation would be closer to the

Table 8. Percentages of errors in the machine translations of the six NMT application systems.

	Percentages						
Type of Errors	Google	DeepL	Amazon	Baidu	Volctrans	IFYTEK	The average percentage of errors in the machine translations of the six NMT application systems
1. Lexical error	69.94%	70.84%	70.75%	67.19%	69.09%	67.84%	69.28%
1.1 Errors in Terminology	63.72%	64.44%	63.91%	61.41%	63.28%	62.09%	63.14%
1.2 Wrong part-of-speech or tense	4.56%	4.53%	5%	4.09%	4.13%	4.16%	4.41%
1.3 Improper use of words or incorrect collocation	1.7%	1.9%	1.8%	1.7%	1.7%	1.6%	1.73%
2. Syntactic error	25.53%	28%	23.69%	19.72%	25.13%	18.56%	23.44%
2.1 Wrong segmentation of a long difficult sentence	2.2%	1.8%	1.7%	2.4%	2%	1.3%	1.9%
2.2 Mistranslation of non-subject sentences	13.56%	14.44%	12.69%	8.78%	13.81%	9.03%	12.05%
2.3 Confusion in sentence structure	9.81%	11.72%	9.31%	8.5%	9.34%	8.25%	9.49%
3. Other errors	19.19%	17.09%	14.5%	18.97%	18.66%	12.97%	16.9%
3.1 Improper Cohesion and Coherence	9.31%	9.22%	8.69%	8.09%	9%	7.5%	8.64%
3.2 Errors in capital and small letters	8.1%	6.3%	4.5%	8.5%	8.1%	4.1%	6.6%
3.3 Omissions	1.8%	1.6%	1.3%	2.4%	1.6%	1.3%	1.67%

ST and more accurate if “the Boao Forum for Asia” was included in the machine translation databases.

(2) Wrong part of speech or tense. ST: “阿拉伯谚语说’金字塔是一块块石头垒成的’。” IFLYTEK TT: An Arab proverb says, “A pyramid is made of stones.” Amazon TT: The Arabic proverb says, “A pyramid is a block of stone.” The ST does not mean one pyramid or one stone. Therefore, the part-of-speeches are used incorrectly in the translations. The plurals are better, that is, “Pyramids were built by piling one stone block upon another,” which can reflect the hardships of building pyramids.

(3) Improper use of words or collocation errors. Polysemous words are frequently used in Chinese political documents to express specific meanings in certain contexts. For example, ST: “要坚持系统治理、依法治理、综合治理、源头治理 . . . . . .” DeepL TT: “To adhere to systemic governance, governance by law, comprehensive governance, governance at source . . . .” Google, DeepL, and IFLYTEK used “governance” indiscriminately for the polysemous word “治理.” None of them carried out interpretative translations as the official translation, leading to the identification error in the polysemous words. This is noteworthy in English translations. It is also one of the difficulties faced in machine translations. To take another example, ST: “健全完善立体化社会治安防控体系 . . . . . ” Amazon TT: “improve and improve the three-dimensional social security prevention and control system . . . .” The problem with Amazon’s translation is “improve and improve,” that is, repetition of words. Therefore, the score of the translation is the lowest.

Syntactic Error

The error rate of syntactic errors in the machine translations is much lower than that of lexical errors, accounting for 23.44% of the total number of sentences. English is characterized by hypotaxis, which is realized mainly through syntax, the means to organize individual words into sentences. Major syntactic errors may result in disorder and ambiguity of sentences. Therefore, the analysis of syntactic errors and the study of syntactic formalization have been major projects in machine translation (Koponen 2010).

(1) Errors in the segmentation of long difficult sentences. ST: “要把人民健康放在优先发展的战略地位, 以普及健康生活、优化健康服务、完善健康保障、建设健康环境、发展健康产业为重点, 加快推进健康中国建设, 努力全方位、全周期保障人民健康, 为实现‘两个一百年’奋斗目标、实现中华民族伟大复兴的中国梦打下坚实健康基础.” Google TT: People’s health should be given priority to the

strategic position of development, focusing on popularizing healthy life, optimizing health services, improving health protection, building a healthy environment, and developing healthy industries, accelerating the construction of a healthy China, and striving to ensure people's health in an all-round and full-cycle manner, laying a solid and healthy foundation for realizing the "two centenary goals" and realizing the Chinese dream of the great rejuvenation of the Chinese nation. The major problem with Google Translation is that there is no sentence segmentation. The translation pursues unduly mechanical formal equivalence to the ST, failing to follow the English way of expression.

(2) Error in non-subject sentences. ST: "空气、水、土壤、蓝天等自然资源用之不觉、失之难续." Google TT: Air, water, soil, blue sky, and other natural resources are unknowingly used and difficult to sustain. The subject of "用之不觉、失之难续" should be a person. Without a subject, the sentence should be translated in the way translates subject-prominent language, that is, with "we" as the subject, to highlight the fact that it is people rather than resources that cannot survive without natural resources. Both Google and DeepL translations fail to handle the non-subject sentence correctly.

(3) Disordered structural relationship. ST: "大道至简, 实干为要." Amazon TT: The road is simple, and practical work is essential. Without giving the implied logical relationship, the translation treats the two parts as parallel structures and expresses an irrelevant meaning. For the sake of correctness, the translation should be "Great visions can be realized only through actions."

Other Types of Error

There are also other types of errors in the translations of political documents, which account for 16.9% of the total errors in machine translations. The low error rate is highly correlated with the occurrence of the words themselves in the sentence. The followings are examples of errors in the machine translations.

(1) Improper cohesion and coherence. ST: "新世纪以来 . . . . . " Amazon TT and DeepL TT: "Since the new century . . . ." Without "the beginning of," both translations are incohesive and incoherent in the context.

(2) Errors in capital and small letters. ST: "坚持亲、诚、惠、容的周边外交理念" DeepL TT: Adhere to the peripheral diplomacy concept of affinity, sincerity, benefit, and tolerance. This headline is not capitalized in the machine translation as it is unidentified.

(3) Omission. ST: “中国同周边国家贸易额由1000多亿美元增至1.3万亿美元 . . . . . ” “Trade” is not described in the translations. “Trillion US dollars,” instead of “trillion-worth,” is expressed in the translations of the six NMT application systems. The translation of “额” in the Chinese context is omitted. It is known from the machine translations of sentences scored “0” that Baidu fails to translate some sentences, e.g., “这是习近平在十八届中央政治局常委同中外记者见面时讲话的主要部分,” “这是习近平在主持十八届中央政治局第一次集体学习时的讲话.”

Scheme for Improving the Translation Quality of NMT Application Systems

It is a fact that the mistranslation of specific words is more prominent in the translations of NMT application systems. Once a certain word is mistranslated, the translation of the whole paragraph or even the whole text will deviate greatly from the intended meaning of the source text. This would have a great impact on the overall quality of the translation. To resolve this prominent problem, this paper proposes a “Cue Lexicon+” model that integrates “machine translation and translation memory.”

A high-quality Chinese-English translation memory (lexicon) is introduced into the NMT application systems to further examine and proofread certain words such as proper nouns, terms, etc. in the translation results. By doing so, the standard translation of words will be identified and matched, thus improving the translation quality of sentences and passages.

Building a “Cue Lexicon”

To generate a high-quality Chinese-English translation memory (lexicon), a “Cue Lexicon” of political documents that includes the political Chinese-English texts over the past 20 years has been built, with a total of 10 million characters having been included by far. The method is as follows:

(1) Content selection. The selected materials are all Chinese-English text materials published by authoritative Chinese institutions to ensure the high quality of the Chinese and English materials, i.e., Chinese political documents since 2000, including speeches and addresses by state leaders of China at international events, news from the website of the Ministry of Foreign Affairs, and reports on the work of the Chinese government.
(2) Generating a corpus and extracting terms. The bilingual texts collected were aligned to make a parallel corpus using the Aligner (See Figure 9). At the same time, specific words and corresponding translations were extracted from the glossary of political news based on word frequency to

Corpus of political documents				
No.	**Language Pair**	**Source**	**Target**	**Remarks**
1	CN-->EN	中国将努力构建总体稳定、均衡发展的大国关系框架，积极同美国发展新型大国关系，同俄罗斯发展全面战略协作伙伴关系，同欧洲发展和平、增长、改革、文明伙伴关系，同金砖国家发展团结合作的伙伴关系。	China will promote efforts to put in place a framework of major country relations featuring general stability and balanced growth. We will strive to build a new model of major-country relations with the United States, a comprehensive strategic partnership of coordination with Russia, a partnership for peace, growth, reform and civilization with Europe, and a partnership of unity and cooperation with other BRICS countries.	
2	CN-->EN	中国将继续坚持正确义利观，深化同发展中国家务实合作，实现同呼吸、共命运、齐发展。	China will continue to uphold justice and friendship and pursue shared interests, and boost pragmatic cooperation with other developing countries to achieve common development.	
3	CN-->EN	中国将按照亲诚惠容理念同周边国家深化互利合作，秉持真实亲诚对非政策理念同非洲国家共谋发展，推动中拉全面合作伙伴关系实现新发展。	We will further enhance mutually beneficial cooperation with our neighbors based on friendship, good faith, mutual benefit, and inclusiveness. We will pursue common development with African countries in a spirit of sincerity, affinity and good faith and with a result-oriented approach. And we will elevate our comprehensive cooperative partnership with Latin America to a higher level.	
4	CN-->EN	第四，中国支持多边主义的决心不会改变。	Fourth, China remains unchanged in its commitment to multilateralism.	
……	……	……	……	
……	……	……	……	

Figure 9. Corpus of political documents.

Cue lexicon						
No.	**Language Pair**	**Source**	**Target**	**Word Length**	**Word Type**	**Remarks**
1	CN-->EN	中国特色社会主义	Socialism with Chinese Characteristics	12	Technical term	
2	CN-->EN	中国特色社会主义制度	the socialist system with Chinese characteristics	10	Technical term	
3	CN-->EN	社会主义初级阶段	the primary stage of socialism	8	Technical term	
4	CN-->EN	社会公平保障体系	the system for guaranteeing social equity	8	Technical term	
5	CN-->EN	拒腐防变	combat corruption and prevent degeneracy	4	Technical term	
6	CN-->EN	软骨病	lack of backbone	3	Technical term	
7	CN-->EN	十八大	the 18th National Congress	3	General term	
8	CN-->EN	科学发展观	the Scientific Outlook on Development	5	General term	
9	CN-->EN	上层建筑	the superstructure	4	General term	
10	CN-->EN	官僚主义	bureaucratism	4	General term	
11	CN-->EN	马克思	Karl Marx	3	Person name	
12	CN-->EN	邓小平	Deng Xiaoping	3	Person name	
……	……	……	……	……	……	

Figure 10. Cue lexicon.

make a cue lexicon (See Figure 10). In addition, categorical attributes were added to the cue lexicon, namely, general terms, technical terms, organization names, place names, and person names, for finer management and expansion of the cue lexicon in the future.

Implementation of the Scheme

This study proposes the "Cue Lexicon+" model (See Figure 11) to perform three functions based on the above design: (1) Importing the cue lexicon to match specific words in the text to be translated automatically by reference to the lexicon. The specific words are tagged (i.e., labeled with hidden tags) so that they are automatically regarded as phrases in the process of machine translation, ensuring the integrity and specificity of phrases; (2) Connecting and fusing the six NMT application systems to translate the tagged texts automatically; (3) Making intelligent comparisons

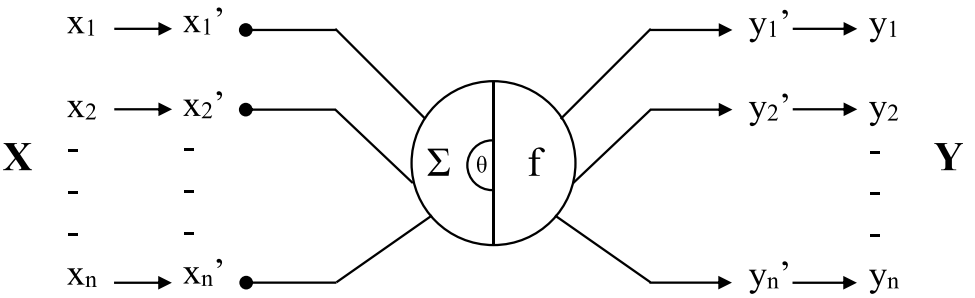

Figure 11. "Cue Lexicon+" model.

among specific lexical expressions in the translations of the NMT application systems.

They would be replaced and updated if the expressions are inconsistent with those in the lexicon, and a final translation will be generated at last. An application named "NMT + Lexicon Intelligent Translation Assistant" (See Figure 12) has been constructed based on this model. The process flow chart is shown in Figure 13.

In this model, X denotes ST; x_1, x_2, x_n denote terms in the cue lexicon; x_1,' x_2' and x_n' denote tagged terms; Σ and f denote pre- and post-translation, respectively; θ denotes the comparison of the machine-translated term with the reference translation in the cue lexicon; y_1,' y_2,' y_n' denote the

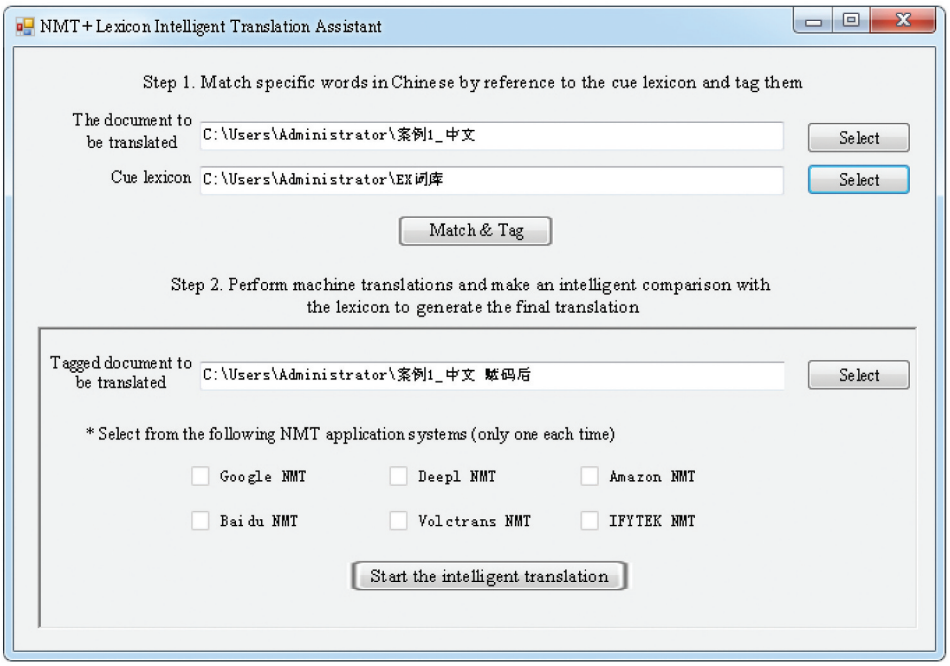

Figure 12. NMT+ Lexicon intelligent translation assistant.

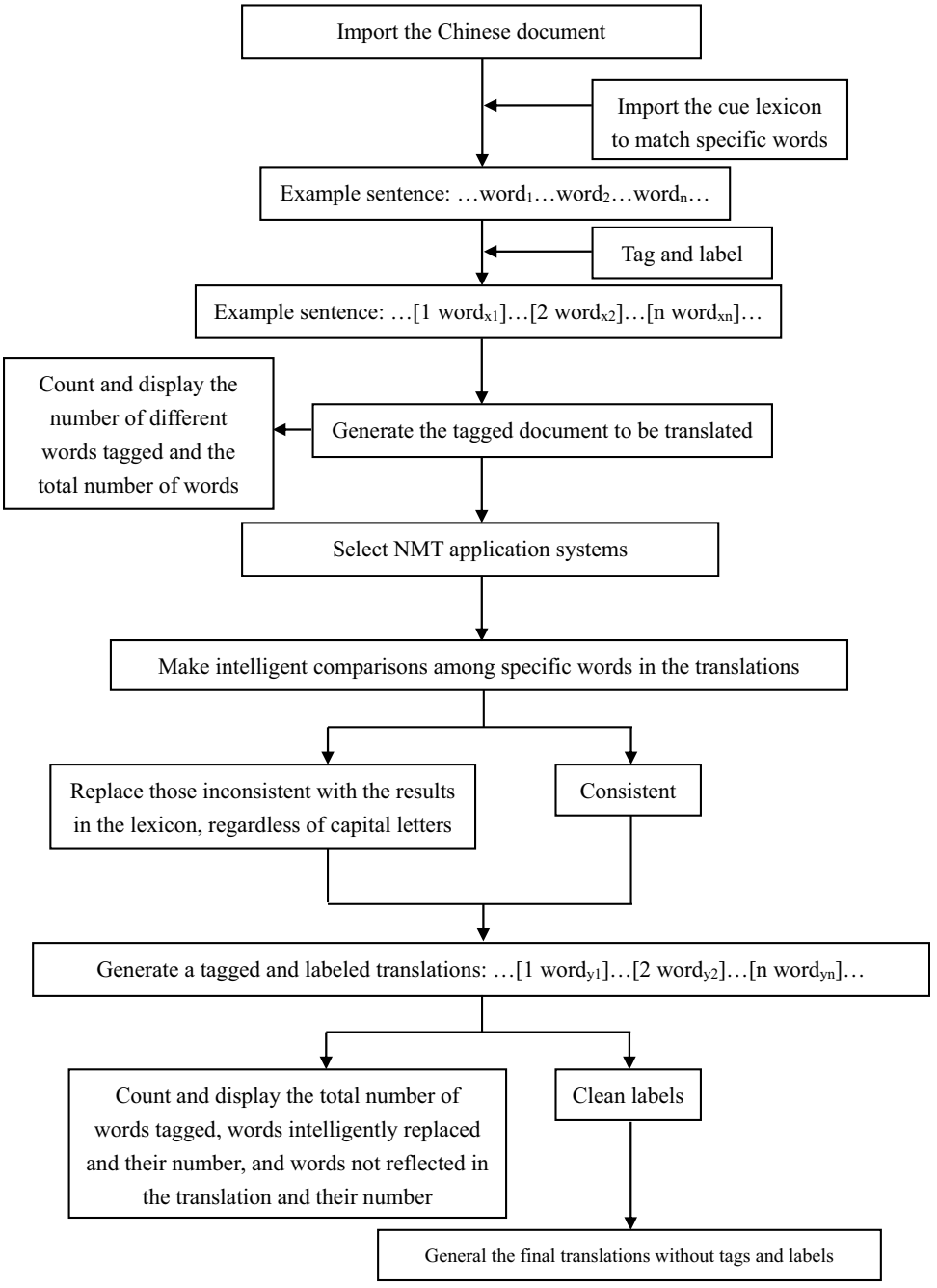

Figure 13. Process flow chart of "Cue Lexicon+.

translated terms (with labels); y_1, y_2 and y_n denote the terms after cleaning the labels; Y denotes TT.

The proposed method is expressed as follows: In the first stage, a Chinese text is uploaded. Cue lexicon is imported. A matching process is conducted

Table 9. Number of words intelligently replaced and updated by "NMT + Lexicon intelligent translation assistant" in the six TT_{new}s.

Total number of words intelligently updated					
Google	DeepL	Amazon	Baidu	Volctrans	IFYTEK
TT_{new}	TT_{new}	TT_{new}	TT_{new}	TT_{new}	TT_{new}
5389	5629	5574	5491	5466	5358

Note: Some words appeared repeatedly.

Table 10. Comparison among BLEU scores of six TT_{new}s and TTs of case 1.

Research object: Case 1 (Xi Jinping: The Governance of China I)						
	BLEU Score					
Description of Translation	Google	DeepL	Amazon	Baidu	Volctrans	IFYTEK
TT_{new} generated after intelligent processing	0.3332	0.3259	0.3199	0.3405	0.3343	0.348
Original TT	0.2099	0.1946	0.1934	0.2148	0.2091	0.2284

between reference and cue lexicon to find matched words and they are tagged. A tagged document is generated to be translated. All different tagged words are counted and shown concerning tagged words. In the second stage, all six translation applications are utilized. Comparisons of translated words are conducted intelligently resulting in tagged and not tagged translations. The number of all tagged words, the number of words replaced intelligently, and the number of words not matched with proper translations are counted and shown in the output.

Test Checkout

By reference to the official Chinese-English texts of Cases 1 and 2, 3,298 specific words were extracted from the glossary of political news based on word frequency to make a Chinese-English cue lexicon. Then, the Chinese texts of Cases 1 and 2 were imported into the "NMT + Lexicon Intelligent Translation Assistant," and the cue lexicon was imported as well to give matched words to be tagged. The texts were subsequently translated with the six NMT application systems to generate final translations (TT_{new}s). The following is the number of words replaced and updated intelligently by the software in the process of translation (See Table 9):

After further BLEU and NIST evaluation of the six TT_ns of Cases 1 and 2, the following results were obtained.

Table 11. Comparison among NIST scores of six TT_{new}s and TTs of Case 1.

Research object: Case 1 (Xi Jinping: The Governance of China I)						
	NIST Score					
Description of Translation	Google	DeepL	Amazon	Baidu	Volctrans	IFYTEK
TT_{new} generated after intelligent processing	7.257	7.1605	7.0408	7.3942	7.2783	7.5062
Original TT	5.8817	5.6822	5.6039	5.9728	5.8621	6.1809

Table 12. Comparison among BLEU scores of six TT_{new}s and TTs of Case 2.

Research object: Case 2 (Xi Jinping: The Governance of China II)						
	BLEU Score					
Description of Translation	Google	DeepL	Amazon	Baidu	Volctrans	IFYTEK
TT_{new} generated after intelligent processing	0.2768	0.2679	0.2653	0.2813	0.2757	0.2979
Original TT	0.2015	0.1878	0.1873	0.2064	0.2005	0.2233

Table 13. Comparison among NIST scores of six TT_{new}s and TTs of Case 2.

Research object: Case 2 (Xi Jinping: The Governance of China II)						
	NIST Score					
Description of Translation	Google	DeepL	Amazon	Baidu	Volctrans	IFYTEK
TT_{new} generated after intelligent processing	6.6459	6.52	6.4459	6.7335	6.6195	6.9871
Original TT	5.695	5.52	5.4627	5.7752	5.6656	6.02

According to the results of technical evaluations, the scores of the six TT_{new}s of Cases 1 and 2 have increased significantly (See Table 10, Table 11, Table 12 and Table 13). The overall quality of the translations processed by "NMT+ Lexicon Intelligent Translation Assistant" has been greatly improved has been shown.

Conclusion

This paper made a comparative study on the performances of six mainstream NMT application systems in the Chinese-English translations of political documents by employing technical and manual evaluations. After comparing and analyzing the translations of the six NMT application systems with the standard translations, this paper concludes problems in the machine translations and builds targeting prominent problems in the "Cue Lexicon+" model and a method called the "NMT+ Lexicon Intelligent Translation Assistant" is proposed, which can greatly resolve the problem of the mistranslation of specific words in the English translation generated by NMT application systems. The results of BLEU, NIST, and manual evaluations of TTs translated with the six NMT application systems were found to be completely consistent. Therefore, TT_{new} generated after intelligent processing resulted in better translation quality.

The research results showed that after the proposed method, the overall quality of machine translation had a qualitative breakthrough. Found that IFLYTEK had the best performance among the six NMT application systems in actual communication.

The research findings may give readers a reference in selecting a machine translation system for political documents and provide a research basis for improving the translation performance of NMT application systems. In

addition, the research proposes a way to build an online corpus platform (http://miaohua.021misn.com) on which the corpus and cue lexicon of political documents are available. Researchers and developers of the NMT application system are welcome to use them for reference.

Disclosure Statement

No potential conflict of interest was reported by the authors.

Funding

This study was supported by Postgraduate Education Reform and Quality Improvement Project of Henan Province (No. YJS2021AL057), and Educational Research Project of China National Committee for Translation & Interpreting Education (No. MTIJZW202130).

References

Catford, J. C. 1978. *A linguistic theory of translation*, Vol. 79. Oxford: Oxford University Press.

Duh, K. 2008. Ranking vs. regression in machine translation evaluation[C]//Proceedings of the Third Workshop on Statistical Machine Translation, Columbus, Ohio, USA, 191–94.

Feng, Y., W. Xie, S. Gu, C. Shao, W. Zhang, Z. Yang, and D. Yu. 2020. Modeling fluency and faithfulness for diverse neural machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (01): 59–66.

Ghorbani, B., O. Firat, M. Freitag. 2021. Scaling laws for neural machine translation. *arXiv Preprint arXiv: 210907740*

Guzmán, F., S. Joty, L. Màrquez, and P. Nakov. 2017. Machine translation evaluation with neural networks. *Computer Speech & Language* 45:180–200. doi:10.1016/j.csl.2016.12.005.

Koponen, M. 2010. Assessing machine translation quality with error analysis. *Electronic proceeding of the KaTu symposium on translation and interpreting studies*.

Luo, J. M., and M. Li. 2012. Error analysis of machine translation. *Chinese Translators Journal* 5:84–89.

Mathur, N., T. Baldwin, and T. Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. doi:10.18653/v1/2020.acl-main.448.

Reiss, K., and E. F. Rhodes. 2014. *Translation criticism–the potentials and limitations: Categories and criteria for translation quality assessment[M]*. Routledge.

Shapran, N. V., S. V. Novoseletska, E. K. Koliada, T. I. Musiichuk, and K. V. Simak. 2021. Communicative-functional components of discourse. *Linguistics & Culture Review* 5 (S4):1325–39. doi:10.21744/lingcure.v5nS4.1784.

Sun, Y., and V. Kumar. 2022. Analysis of Chinese machine translation training based on deep learning technology. *Computational Intelligence and Neuroscience* 2022:1–14. doi:10.1155/2022/6502831.

Wu, L., X. Pan, Z. Lin. 2020. The volctrans machine translation system for wmt20. *Proceedings of the 5th Conference on Machine Translation (WMT)*, 2020 November 19–20, 305–312.

Wu, Y., M. Schuster, Z. Chen. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv Preprint arXiv: 160908144*

Yulianto, A., and R. Supriatnaningsih. 2021. Google translate vs. DeepL: A quantitative evaluation of close-language pair translation (French to English). *AJELP: Asian Journal of English Language and Pedagogy* 9 (2):109–27.

Zhou, J., Y. Cao, X. Wang, P. Li, and W. Xu. 2016. Deep recurrent models with fast-forward connections for neural machine translation. *Transactions of the Association for Computational Linguistics* 4:371–83. doi:10.1162/tacl_a_00105.