International Journal of Frontiers in Engineering Technology
ISSN 2706-655X Vol.1, Issue 1: 30-37, DOI: 10.25236/IJFET.2019.010103

Analysis and research on lexical errors in machine translation in Chinese and Korean translation

Lina Leng[1,*], Guangyong Shan[2]

[1] Northwest University of Political Science and Law, Xian 710063, China
[2] Harbin Institute of Technology, Weihai 264209, China
[*]Corresponding author e-mail:42142963@qq.com

ABSTRACT. *In the current Korean language major learning, there are not a few students who rely on machine translation to complete the learning, Especially for beginner and intermediate students. Due to their lack of Korean proficiency, they cannot make a correct judgment on the quality of the translation, and the current machine translation quality has not reached a high level. Therefore, it is particularly important to analyze the errors in machine translation to improve the quality of the output text of machine translation and to make appropriate post-editing.*

KEYWORDS: *Machine Translation，Vocabulary，Korean, translation practice*

1. Introduction

Human society is a complex whole made up of various cultures. The influence and integration of different cultures have driven the development of human society. I don't know when the main body of cultural communication became language. Therefore, translation between different languages has become the key to cultural communication. Especially in the information age, translation plays a more important role in information exchange.

In recent years, China and Korea have become increasingly active in exchanges in the fields of politics, economy and culture between the two countries. In the past, in order to facilitate cultural exchange, text information of other languages was often obtained through a translator with expert knowledge. With the development of traffic and communication, people are pursuing high speed in all aspects of life in the global era. Unlike the past, people use machine translator in everyday life to quickly access information from other languages and quickly obtain results in the desired language in the field of translation.

It is obvious that high-quality translation services have a positive impact on the development of social economy, but traditional human translation is far from enough to meet the growing demand for multi-language translation. With the wide application of computer, the research on machine translation has attracted more and more attention. With the progress of The Times, the technology of machine translation is becoming more and more perfect, and various kinds of machine translation software are widely used by the public. In the current Korean language major learning, there are not a few students who rely on machine translation to complete the learning, Especially for beginner and intermediate students. Due to their lack of Korean proficiency, they cannot make a correct judgment on the quality of the translation, and the current machine translation quality has not reached a high level.

2. Theoretical background

2.1 Machine Translation

In X.G.Liu[1], Machine translation is a new science that USES computers to translate through translation systems. Machine translation by means of a computer, a document in one language is translated into a document in another language. Machine translation involves linguistics, computer science, mathematics and other professional fields.

From the perspective of the development history of machine translation, it is actually the earliest research area of linguistics, including speech recognition and linguistics. Over the last few decades, in order to realize the dream of machine translation, represented by IBM, Google, Microsoft's foreign scientific research institutions and companies have set up machine translation team, specializing in intelligent translation study, to break the Babel of languages, but a breakthrough occurred in 2014, the field of machine translation has earth-shaking changes, occurs during this event, is based on machine learning neural network as the foundation of machine translation, in full beyond previously based on the statistical model of statistical machine translation (SMT), and quickly became the mainstream standard of online translation system.

2.2 Principles and methods of machine translation

The development of machine translation can be seen from two aspects: the language processing method of machine translation and the basic technology of machine translation. With the development of basic technology, the language processing method of machine translation has developed from simple substitution vocabulary to the level of using intermediate language. The machine translation approach goes through four phases. In Z.G.Ling[2], In the first stage of word translation, word to word translation is carried out. The second stage is phrase translation, grammar analysis and phrase translation; The third stage of semantic analysis, through the analysis of words and sentences to carry out semantic analysis; In the fourth stage, intermediate language is used for analysis. The above language processing methods provide a theoretical background for the research of analyzing wrong words by machine translation.

In BaiDu[3]At present, the two most important methods of machine translation are rule method and statistics method. Rule method, according to the language rules of the text analysis, and then with the help of computer programs for translation. Most commercial machine translation systems adopt the rule approach. The operation of regular machine translation system is realized through three consecutive stages: analysis, transformation and generation. Direct translation is simple word-to-word translation. Translation refers to the lexical, syntactic and semantic information of the original text. Because the range of information sources is too broad, there are too many grammatical rules and there are contradictions and conflicts among them, conversion and translation are complicated and easy to make mistakes. Statistical machine translation (SMT), through statistical analysis of a large number of parallel corpora, constructs a statistical translation model, such as vocabulary comparison or language model, and then USES this model for translation. Generally, the entry with the highest probability in statistics is selected as the translation. The probability algorithm is based on bayes' theorem. Suppose you want to translate an English sentence A into Chinese, and all Chinese sentences B are possible or non-possible potential translations of A. Pr(A) is the probability of something like A, and Pr(B, B, B, 0 A) is the probability of A being translated into B. Finding the maximum value of two parameters can narrow down the scope of sentence and corresponding translation retrieval, so as to find the most suitable translation. SMT is divided into two types according to the degree of text analysis: word-based SMT and phrase-based SMT, the latter of which is currently in common use, such as that used by Google. The translated text is automatically divided into fixed-length word sequences, and then the word sequences are statistically analyzed in the corpus to find the translation with the highest corresponding probability.

There are many theories about the principles and standards of translation quality. This thesis focuses on the study of vocabulary and reconstructs the standards for determining the quality of vocabulary translation based on the previous translation standards. That is, whether the translation can fully and accurately express the meaning of the original words, whether the translation is consistent with the original text, whether the translation can faithfully express the emotions of the characters in the original text, and whether the translation can translate culture-related words according to the characteristics of The Times of the original text.

3. Incorrect lexical analysis in machine translation

3.1 Research object and research method

The text of this paper is the text learned in the reading class, all of which are selected from the short stories about daily life. This paper focuses on the vocabulary in the sentences of novels and analyzes the errors in machine translation. The reason is that vocabulary is the collection of words, is an important part of sentence components. Part of speech is classified according to the grammatical properties of words. The classification of a word in a sentence is the grammatical relationship between the word and other words. Therefore, if the problem of machine translation is solved by starting from the vocabulary of articles, the error rate of literary works and other machine translation will be greatly reduced.

In this paper, 10 representative sentences were selected from the reading materials and machine translation was carried out by Naver Papago and Google, two online platforms with the highest evaluation rates among Korean majors. On the basis of the obtained translation, the lexical errors frequently found in the two kinds of online machine translation software are viewed in the form of statistical data. Finally, the errors are analyzed by type through specific patterns. When analyzing the error part, "original text-ST", "google-MT-G", "papago-MT-P", "human-transaction-TT" will be used for annotation. For the wrong words, a comparative analysis will be made according to the semantics marked in the standard mandarin Chinese dictionary.

3.2 Analysis of lexical errors in machine translation

(1) General vocabulary error. Through the observation of lexical errors in machine translation, it is mainly embodied in the substitution error and omission error. It is analyzed from three perspectives, that is, two words with different meanings, words that enlarge the meaning of the original text and words that reduce the meaning of the original text.

Table 1 General vocabulary error

NO	ST	MT-G	MT-P	TT
①	Tao ying rides the bus alone and often doesn't buy a ticket	종종	항상	항상
②	Tao ying is anxious, want to solve this matter quickly, her child is waiting for her	일	일	문제
③	His hair was as shaggy and dull as hay	어둡다	빛도 없다	어둡고 윤기가 없다
④	"Go in! Don't stand in the doorway! It's not a train. A stop from Beijing to baoding, soon to the station..." 'cried the conductor impatiently shout	매표원… （missing）	매표원… （missing）	매표원… 외쳤다 （missing）

In example①, The word ST's "often" in the example is used to describe a person's consistent behavior whenever it comes to things. According to the context of the article, is to decorate the hostess fair when the bus does not buy this behavior. Google translate into "종 종". Papago translated as "항 상". In Google's translation, words that are not consistent with the original text appear, leading to the deviation of the meaning of the original text. In example②, The word "things" in ST has been translated to mean something else. In the original text, the word "matter" in "solve this matter" is analyzed in a broad sense corresponding to the word "matter" in Korean. However, the narrow sense only corresponds to one of several meanings. In example③, The word ST's "dark and dull" means dull and dull in Chinese. However, in both Google and Papago's machine translation, there are different degrees of absence, which can be seen in comparison with the original text, reducing the original meaning of words. In example④, The word ST's "shout" is completely missing from Google and Papago's machine translation, as the lack of content makes readers unable to properly understand the meaning of the original text. Thus it can be seen that machine translation violates the principle of word generality in some cases.

(2) Culture-related vocabulary error. Based on the results of machine-translated lexical errors, Google and Papago translators have seen a number of idiomatic lexical errors. The following will be analyzed and illustrated from three aspects: idiomatic vocabulary, culture-specific vocabulary and inherent noun vocabulary.

Table 2 Culture-related vocabulary error

NO	ST	MT-G	MT-P	TT
⑤	Working with oil all day long, nails are shiny and shinning like seashells	기름을 다루다	얼굴에 기름칠을 했다	기름과 밀가루를 가까이 하다
⑥	Between xiaoye's round head and the standard line for buying tickets, there were tao ying's long and beautiful fingers	표준 라인	표준선	매표선
⑦	The admission fee is 50 cents. Now the temple is worth so much. The ticket is from Lao zhang on the red case	Lan Zhang	장 씨	라오장

In example⑤, The word ST's "to deal with" is a common expression in Chinese. The scene of the heroine's fingernails glinting like seashells is reflected in the original text, in which she is a baker who has to come into contact with flour every day. But the Google translator translates directly as "dozen," and the Papago translator translates as "shangyou." The translation of the two translators is quite different from the original. It can be seen that in terms of the translation of idioms, machine translation is still unable to meet the requirements of the meaning expressed in the original text. In example⑥, The conductor measures the child's height according to the specified height and decides whether he needs to buy a ticket. Cultural differences and lexical meanings should be taken into account when translating vocabulary in a specific context. Both Papago and Google's translation of "standard line" does not match the context and meaning of the original text, leaving readers unable to understand the meaning of the sentence. In example⑦, The word ST's "Lao zhang" is short for Chinese surnames. In Chinese expressions, people's names are often omitted to address each other, while similar expressions in Korean are also found, but the omitted parts are not the same. As a result, neither Papago nor Google was able to translate it accurately and properly.

(3) Other word errors. In addition to the lexical errors mentioned above, this paper also finds out that there is a problem of spacing words errors in machine translation when analyzing the translation results, and the frequency of spacing words errors is also very high.

Table 3 Other word error

NO	ST	MT-G	MT-P	TT
⑧	Tao ying rides the bus alone and often doesn't buy a ticket	혼자있을 때	혼자서 버스를 탈 때	혼자 버스를 타면
⑨	"Mom is coming," tao ying answer with a loud voice	엄마가가고있어.어서	엄마다 와요.그냥 와	그래, 그래, 금방 갈게
⑩	"Who said no tickets?" Asked the young man in red, tilting his head.	누가 표를사지 말라고?	누가 투표하지 말라고 했지?	누가 표가 필요없다고 했냐?

As can be seen from the translated results of ST⑧, ⑨and ⑩, there are all problems of wrong writing methods, among which Google has many mistakes. When writing an article, divide a sentence from the preceding sentence according to its grammar. After machine translation, the parts that need to be separated in the above examples violate the rules of division, so they should be regarded as ungrammatical sentence.

In ancient literature, there is no use of fractional writing at all, but the current orthography stipulates that fractional writing should be carried out with words as units, but auxiliary words should be appended after the preceding words. Since the auxiliary words in Korean are less independent as words, such a measure is naturally adopted. For example, the spelling of compound words in English is not uniform.

For example, "bathroom, high chair, ape-man" and other words are spelled in both coincidences and separations, and some even add conjunctions in the middle. The Korean orthography does not have these inconveniences after all, because Korean compounds are always co-written. In fact, though, the most difficult part of Korean orthography is the division, which causes the most confusion in books, newspapers and other publications. Such lexical errors are common in machine translation. Therefore, in the use of various machine translation, attention should be paid to ensure that the correct or not.

4. Conclusion

Machine translation has advantages of getting results quickly, but has many limitations. The study tried to translate machine translation by choosing between Google Translator and Naver Papago Translator, which are representative of machine translation based on neural network technology. But, this study only uses the Google and Naver Papago translator as a representative means of machine translation, and does not focus on comparing the two. Instead, through the attempts of the two translators, The study wants to extract the errors of the vocabulary in the overall machine translation as a target for analyzing the literary works and conduct a detailed analysis through statistical numerical values. Then, based on the previous research presented in Section 3. the limitations of machine translation are examined in detail. Then, based on the experimental results of this study, the limitations identified in the previous research are confirmed and a new error tendency is presented. Furthermore, we aim to propose the role of human translators to complement the limitations of machine translation presented in this study.

Through Papago's and Google's online translation experiments, it can be concluded that machine translation has the highest error rate when dealing with words in articles: substitution errors and omissions. Culture-related vocabulary errors are mainly reflected in the lack of understanding of vocabulary, the analysis of the meaning before and after the article and the lack of understanding of the emotional characters in the article. In addition, there are some other vocabulary errors. To sum up, we can conclude that no matter what level of machine translation development, there are still some limitations from a certain point of view, and can not completely replace human translation. For Korean language majors, while using various types of machine translation for learning, they need to realize that machine translation is only an auxiliary means, which can be referred to according to the results of translation, but cannot be taken as the final answer. Many people also have a misunderstanding about machine translation, thinking that the result of translation is too big to be used with confidence. In H.B.Gao[4], Machine translation USES the principle of linguistics, the machine automatically recognizes the grammar, calls the stored thesaurus, and automatically translates the corresponding words. However, due to the changes or irregularities in grammar, morphology and syntax, errors are inevitable. In this paper through the machine translation text summary and induction, finishing some vocabulary mistakes in current machine translation, due to the limitations of the research object, in the future research, more research will be based on the analysis object, to find a more complete error form, and carries on the induction more rigorous study, strive for more meaningful results.

References

[1] X.Ging. Liu (2015). Research and review on machine translation in China. Problems and countermeasures[M], Beijing Science Press, p.10-12.

[2] C.Ken. Lee (2014). Improvement of Translation Processing Performance of Neural Network Based Machine Translation with Interest Inference and Target Word Recommendation. Yonsei University, p.23-47.

[3] BaiDu. https://blog.csdn.net.

[4] H.Bo. Gao (2019). R esearch on the A pplication of M achine Translation in Intensive R eading of C ollege English. Journal of HeiHe University, vol.58, no.4, p.120-121.

[5] Z.Yun. Chen (2014). Research on the Strategies of Integrating Machine Translation Into College English Teaching, vol.32, no.1, p.74-77.