Segmented Translation Algorithm of Complex Long Sentences Based on Semantic Features

Yu Shi*
Wuhan Donghu University, Wuhan, China

*Corresponding author: cynthia3379-cn@126.com

Abstract. The progress of the times is inseparable from communication. If a country wants to develop well, it must learn from each other. Accurate language translation can better let people understand what they want to express. Therefore, language translation is becoming more and more important in the current social communication. Although there is a lot of research on translation, in many cases there will be inaccurate translations. Therefore, finding an accurate translation method is what many people need. Aiming at the problem that the translation of complex long sentences is prone to errors, this paper proposes a sentence segmentation algorithm, which is a method of dividing the long sentence into multiple independent clauses and then translating it. The segmentation algorithm uses the semantic features of the Concept Hierarchical Network (HNC) theory to segment clauses. The segmentation algorithm is integrated with a rule-based baseline translation system. The BLEU value of the integrated translation system reaches 0.1898, which is higher than that before the integration. The system has increased by 30%. Experimental results prove that the proposed method can effectively improve the effect of patent translation.

Keywords: Semantic Features, Complex Long Sentences, Segmented Translation Algorithms, Concept Hierarchical Network Theory

1. Introduction

1.1 Background and Significance
Today's society has evolved into an information society. With the rapid development of information technology such as computers, software, and networks, mankind has entered the information age [2]. Against this background, people increasingly need to quickly process various language and text information and extract the useful parts from it. In order to meet the needs of human beings, the research field of natural language processing and understanding has emerged [6]. Natural language processing and comprehension is a comprehensive subject integrating cognitive, computer science, linguistics, mathematics and psychology. As a way to explore the intelligent behavior of human understanding of natural language, the research of natural language understanding has made considerable progress in the past two decades [7]. The research content in this field of natural language
processing is very rich, mainly including machine translation, information retrieval, question answering systems, etc. With the popularization of the Web, the application of natural language processing technology has blossomed everywhere, and fruitful results have been achieved in the rapid processing of language and text information [8].

English is the most widely spoken language in the world. According to statistics, a total of 67 countries around the world use English as an official language, more than 400 million people use English as their mother tongue, and more than one billion people use English as a second or foreign language. With the continuous deepening of my country's reform and opening up, my country's exchanges with countries around the world have become more and more frequent. However, the number and quality of existing translators are far from meeting the needs of my country's extensive communication with the world [9]. Therefore, for our country, the study of English-Chinese machine translation is of great value. If the excellent scientific research results in the field of English-Chinese machine translation can meet or partially meet people's requirements, it is likely to bring considerable economic benefits to our country [10].

1.2 Related Work
At this stage, combining traditional rule-based methods with corpus-based methods has become the mainstream of machine translation research. Nakajima et al. used a method of grammaticalized target language phrases to process and combine statistical translation results to form grammatical translation results [1]. Wang et al. proposed a method of regrouping, marking and aligning the translation results using statistical methods. In the Japanese-English patent translation, Jeon J proposed a method of post-editing the results of rule translation using statistical methods [3]. Jergler proposed a phrase alignment method, using the rule method to process translation corpus for statistical translation [4]. The State Intellectual Property Office of China has proposed a hybrid translation method that integrates semantic processing methods, and has achieved good results in Chinese-English translation [5]. These methods are very useful in most translations, but in the translation of complex sentences, there will still be problems such as sentence failure.

1.3 Main Content
In order to solve the problem of wishful errors in the translation of complex long sentences, this article adopts a sentence segmentation algorithm to segment and translate the content of complex long sentences, divide them into simple short sentences, and then combine them. This paper adopts the sampling survey method to translate the selected long sentences, and finally obtains the translation results and compares them with traditional translations. In the end, we get the result: The method in this paper can reduce the occurrence of this situation, and the sentence accuracy rate has increased by 30%.

2. Research Methods of Segmentation Translation Algorithms for Complex Long Sentences Based on Semantic Features

2.1 Segmentation of Complex Long Sentences
A Chinese sentence is a string ending with a period (or a question mark or exclamation mark equivalent to the function of the period). The period is separated by one or more commas (or semicolons). According to statistics in this article, the proportion of Chinese strings ending with a comma alone is more than 65%. Therefore, the sentence segmentation method in this article will use commas as a candidate for segmentation. In this article, the author believes that the semicolon in Chinese is similar in function to the comma, and does not distinguish it.

In this article, CSP represents a string ending with a period, and CSC represents a string ending with a comma. Therefore, a sentence can be expressed as: CSP = CSC, CSC,..., CSC.

The algorithm of sentence segmentation is to determine whether one or more CSCs are a complete clause. If they are, segmentation is performed after the comma of the CSC, and the segmentation mark
"++" is added; if not, skip this Comma, continue processing the next CSC.

This paper uses Concept Hierarchical Network (HNC) theory to guide sentence segmentation, and uses HNC theory's lv hypothesis and characteristic semantic block formation algorithm for segmentation judgment. Assuming a sentence or clause, there is generally only one global core verb (equivalent to the central verb, abbreviated as Eg (global Eigen)). If a CSC contains Eg, it is judged that this CSC is a clause; otherwise, this CSC cannot be regarded as a separate clause. Define Eg candidate set as the set of all possible verbs in CSC. Since there are relatively few adjective predicates in the patent text, this article does not consider adjective predicates for the time being.

Assuming that there are multiple verbs in a CSC, expressed as "...v1...v2...vn..." (n≥0), the method of sentence segmentation can be described as follows:

(a) If n=0, there is no Eg in the CSC, and this CSC cannot be a single clause.
(b) If a certain verb vi (0 <i≤n) in CSC is negated by other language features and cannot be regarded as Eg, then this vi does not enter the candidate set of Eg.
(c) If a certain verb vi (0<i≤n) in CSC can be regarded as Eg by other language features, then this vi enters the candidate set of Eg.
(d) If there are one or more candidates in the candidate set of Eg, this CSC can be a single clause.
(e) If there is no content in Eg's candidate set, this CSC cannot be used as a separate clause.

2.2 Fusion Method of Segmentation Algorithm and Machine Translation System

Due to the differences in expression habits between Chinese and English, the translation results of these clauses need to go through the clause order adjustment module and clause translation synthesis module to be combined into a complete translation result and output. The clause order adjustment module is to adjust the order between the translated English clauses: (a) When translating, first select a core clause from multiple clauses. The core clauses are generally translated into English and adjusted to the position of the first clause, and all other clauses are adjusted after the core clause. (B) Conditional clauses in Chinese, such as LNP such as "in the …", often precede the main sentence, and need to be adjusted after the main sentence after being translated into English.

In addition, Chinese tends to use a series of short sentences to express related content, while English tends to use clauses to express. The clause translation synthesis module realizes this kind of transformation: (a) When translating, the core clause is the backbone and governs the entire sentence. (B) Other clauses, except for the conditional clauses that are processed in the adjustment of clause order, all other clauses are transformed into clause forms. Generally, when there is a temporal relationship or a parallel relationship, it is transformed into a "v + ing" clause, when there is a purpose relationship, it is transformed into a "to + v" clause, and when the tense is passive or past tense, it is transformed into "v + ed" "Clause.

3. Segmentation Translation Algorithm for Complex Long Sentences Based on Semantic Features

This paper selects 1,116 complex long sentences from the patent text as test data, and manually annotates these patent sentences. The annotation content mainly includes sentence segmentation "++" and Eg candidate "[Eg]". This article counts the small sentences contained in these test sentences. Then compare the results translated by traditional translation methods with those translated by the methods of this article. Then judge whether the segmented translation algorithm for complex long sentences based on semantic features can increase the accuracy of translation.

4. Research and Analysis of Segmentation Translation Algorithms for Complex Long Sentences Based on Semantic Features

4.1 Sentence Segmentation Algorithm

Table 1. shows the experimental results of the sentence segmentation algorithm. From Table 1, the author found that the segmentation algorithm has a higher segmentation accuracy rate (90.3%), and its
F value is 85.4%.

Table 1. Comparison of Sentence Segmentation Algorithm and Traditional Translation

| Segment Algorithm | Traditional Translation |
|-------------------|-------------------------|
| Total number of long sentences | 1117 | 1117 |
| Total number of clauses included | 1848 | 1848 |
| Number of clauses given by segmentation algorithm | 1653 | 1653 |
| Correct segmentation | 1494 | 1494 |
| Correct rate | 90.3 | 88.7 |
| Recall rate | 81 | 78 |

Figure 1. Comparison of sentence segmentation algorithm and traditional translation

After error analysis of the segmentation results, it is found that the main problem is: Egp recognition errors caused by the ambiguity of words. For example, the "interval" in Example 1 is recognized as a noun because the second CSC is not divided.

Example 2 This equipment is equipped with a magnet under the bottom of the developing device, which is spaced a certain distance from the bottom of the device to generate a detection signal by detecting the vertical movement of the magnet. The second reason is the recognition error of the characteristic semantic block Egp of large sentences. Egp is identified based on the large sentence expressions that often appear in patents, but the summary of such large sentence expressions is still incomplete. For example, in Example 8, the Egp recognized by the segmentation algorithm is "include", but the first "include" only covers the second comma CSC, and the second "include" covers the last two CSCs.

Example 8. A heat dissipation hole is opened on the wall of the main box. The hot air conveying device includes a heat dissipation fan and a hot air guide device to guide the hot air through the heat dissipation hole to between the main box and the auxiliary box.

4.2 Machine Translation after Fusion

For the evaluation of the translation results, this article uses the internationally universal evaluation index LEU and the universal evaluation tool mteval-v13a. When there is only one standard reference answer sentence, the BLEU-4 value automatically translated by the system is 0.1898. The BLEU-4 value of the rule-based baseline system translation without adding the segmentation algorithm is 0.1457, and the BLEU value of the system after the fusion has increased by 30%. Table 2 shows the evaluation results, where RB-MT represents the baseline system, HYBRID-MT is the fusion system of this article, and GOOGLE is the Google online translation system.
Table 2. Evaluation of Translation Results

| RUNS       | BLEU-4   |
|------------|----------|
| RB-MT      | 0.145 7  |
| HYBRID-MT  | 0.189 8  |
| GOOGLE     | 0.175 0  |

This article compares with the latest level of machine translation in the world: In 2016, the internationally famous NIST machine translation Chinese-English translation large-scale data set (Chinese to English Large Data Track) evaluation, Gale subset (each sentence There is a reference answer translation) best score is 0.1470. The training data and test data used in the NIST machine translation evaluation are all from the news field, and the data scale is much larger than the data scale used in this article. Although the two are not directly comparable, it can be seen that the translation effect of the integrated system has reached or even surpassed the best international news translation level. For the same corpus, the BLEU value of GOOGLE's translation result is 0.1750, and the translation quality of the integrated system is also higher than that of GOOGLE.

5. Conclusions

Aiming at the problem of long sentence translation in patent translation, this article uses sentence segmentation algorithm to process Chinese sentences, divide the long sentence into one or more independent clauses, and use a rule-based translation system to translate the clauses separately. The result of small sentence translation is adjusted by the order of the small sentence and the small sentence translation is synthesized, and finally the result of the large sentence translation is given. The BLEU value of the fusion translation system reached 0.1898, which was 30% higher than the pre-fusion system. This article uses a rule-based sentence segmentation algorithm, but due to the complexity of the patent text, especially its wide distribution of fields and many new words, the next step is to combine statistical algorithms to continue to improve the sentence segmentation algorithm. In addition, the integration of small sentence translation results (arrangement of clause order and clause translation synthesis module) is still relatively simple, and the processing of multiple clauses synthesis is not ideal, and continuous improvement is needed. The segmentation algorithm in this article is relatively independent. The next step will be to try to integrate with statistical machine translation to further improve the effect of patent translation.

References

[1] Nakajima Y, Ptaszynski M, Honma H, et al. A Method for Extraction of Future Reference Sentences Based on Semantic Role Labeling[J]. Ice Transactions on Information & Systems, 2016, E99.D(2):514-524.
[2] Fangzhao L, Changjian W, Xiaohui L, et al. A Composite Model of Wound Segmentation Based on Traditional Methods and Deep Neural Networks[J]. Computational Intelligence & Neuroscience, 2018, 2018:1-12.
[3] Jeon J, Jung J, Kim J, et al. Semantic Reconstruction: Reconstruction of Semantically Segmented 3D Meshes via Volumetric Semantic Fusion[J]. Computer Graphics Forum, 2018, 37(7):25-35.
[4] El-Khoury V, Jergler M, Bayou G A, et al. Fine-granularity semantic video annotation: An approach based on automatic shot level concept detection and object recognition[J]. International Journal of Pervasive Computing & Communications, 2016, 9(3):243-269.
[5] Guan X, Han J, Liu Z, et al. Sentence Similarity Algorithm Based on Fused Bi-Channel Dependency Matching Feature[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2020, 34(07):318-322.
[6] Zhang J, Mu Y, Feng S, et al. Image region annotation based on segmentation and semantic correlation analysis[J]. IET Image Processing, 2018, 12(8):1331-1337.
[7] Tanaya D, Adriani M. Dictionary-based Word Segmentation for Javanese[J]. Procedia Computer Science, 2016, 81:208-213.
[8] Han, Xu, Wang, et al. Sentence segmentation for classical Chinese based on LSTM with radical embedding[J]. The Journal of China Universities of Posts and Telecommunications, 2019, v.26(02):5-12.

[9] Li Xiaohong. An Analysis of Translating Techniques of Long and Complex English Sentences[J]. Overseas English, 2017, 000(006): 150-152.

[10] Fei J, Wen H, Zhang D. Research on infrared image SVM segmentation algorithm for mine drilling and rescue[J]. Journal of Mines, Metals and Fuels, 2017, 65(10):563-569.