Research on Translation Processing of Complex Long Sentences Based on Multi-strategy Analysis Algorithm

Jing Liu*
Wuhan Donghu University, Wuhan, Hubei, 430212, China

*Corresponding author: jingemail2020@126.com

Abstract. Syntax analysis depends on a certain grammatical theory. In formal grammar system, dependency grammar has gradually become the focus of researchers. In this paper, an automatic translation method of English-Chinese terms based on multi-strategy fusion is proposed, which uses the term case base to optimize the translation of terms with low evaluation scores obtained from the translation quality estimation method. Through the levenshtein algorithm, we find a number of terms translation examples with high similarity with the source language of the terms to be optimized, including grammatical semantic features, sentence length, punctuation marks, functional words and context conditions, etc., to segment and simplify complex long sentences and compound generate translations. By designing the same knowledge representation for different languages, the algorithm can be universally used for translation systems of different languages.

Keywords: Multi-strategy analysis, Long sentence segmentation, machine translation

1. Introduction
As human civilization enters the information age, people hope to communicate freely with computers in natural language, and the demand for natural language processing is increasingly urgent. Natural language understanding has become a fascinating and challenging subject. Machine translation methods can be divided into rule-based machine translation and corpus-based data-driven machine translation [1-2]. In many application environments of machine translation system, we often encounter a large number of complex long sentences containing complex clause structures, parenthesis and non-finite verb phrases, in which clauses and phrases often restrict each other and depend on each other, thus forming a complex language phenomenon that there are phrases in clauses and clauses in phrases. All kinds of professional documents are difficult for ordinary people to read and understand because of the existence of professional terms. Some English translations of professional terms contain uncommon words, and some have special expression methods, which makes it difficult to read [3]. However, patent texts have a wide range of fields, many technical terms, long sentences and complex sentence structures, so the effect of patent automatic translation is generally lower than that of ordinary texts.

The research in this paper is aimed at the translation of long sentences in English-Chinese machine translation. As we all know, long sentences are widely used in English. Especially in official texts such
as news, and in fields such as scientific and technological documents and patent documents, long sentences are used more frequently [4]. In this paper, the source language is mainly preprocessed, marked with parts of speech, long sentence segmentation, fragment analysis and translation selection. Finally, a long sentence is divided into small fragments, so a long sentence should be recursion of small sentences, which can simplify the design and implementation of translation system and improve the quality and accuracy of translation.

2. Sentence Segmentation Method

2.1 Semantic Features of Sentence Segmentation

Complex sentences in Chinese are often expressed in a full stop, and conjunctions expressing the relationship of complex sentences are an important feature of CSC segmentation, and such conjunctions usually appear at the beginning of clauses. The system will transform and analyze these subunits into target translation units; Then, according to the consistent results and rules, the system will organize the target language among these subunits according to the semantic relationship, and then output it. For example, for rule-based machine translation system, long sentences will bring great difficulties to syntactic analysis; For machine translation system based on statistics, long sentences will bring a lot of trouble to the step of short intonation. According to a certain standard, the candidate double-word vocabulary is selected, and then some words are expanded left and right on the basis of the candidate double-word vocabulary, so as to obtain the possible multi-word specialized vocabulary set. Usually, the longer a sentence is, the more complex it will be, and at the same time, the search space for sentence analysis will be expanded, which will lead to an increase in ambiguity, thus increasing the complexity of sentence analysis and the probability of error analysis [5].

Terms, as the essence of knowledge in language, are short, concise, clear in meaning and compact in structure. Usually, the length of terms is between 5 and 15 words, which is a special term. Detailed rules can accurately and intuitively describe the grammar and semantic composition of a language, and multi-level rules are convenient for deep understanding and complex structure processing, and different plane transformations are implemented for different sentences, thus effectively solving the problem of long-distance dependence. By specifying a certain search length, all the subsequent decision sequences within this length are scored, and finally the decision sequence with the highest score is used, thus greatly reducing the number of wrong decisions.

2.2 Sentence Segmentation Algorithm

The semantic features related to sentence segmentation are extracted from sentences to formlvsequences (sequences composed of logical concepts and verbs), and thenEgcandidate sets are extracted to judge whether segmentation is possible. With the improvement of rule-based method, the rule-making has increased the proportion of corpus knowledge, that is, general rules and methods are obtained by training and analyzing multiple examples in the corpus, such as the common error-driven learning algorithm [6]. Segmentation and simplification of complex long sentences and compound generation of translations. The features selected by the segmentation algorithm proposed in this paper not only serve the long sentence segmentation task, but also contribute to the translation generation.

In this paper, sentence segmentation algorithm is integrated with a rule-based Chinese-English machine translation system to form a new patent translation system. Figure 1 shows the integrated system architecture.

![Fig.1 Integrated machine translation architecture](image-url)
It can be seen from Figure 1 that three modules are added to the original rule-based translation system: sentence segmentation module, clause order adjustment module and clause translation synthesis module.

Translation norms are not only a yardstick to measure the quality of translation, but also a guideline to guide translation practice. However, patent translation is different from ordinary text translation, which does not require colorful rhetoric and changeable sentence patterns. We need to do some processing on the statistical database. For the double-word items and expanded multi-word items in the statistical database, in addition to the statistics of word frequency, two fields are added: Def and Trans, recording the interpretation and English translation of this item respectively. This shows that the longer the sentence is, the more functional words there are, and the more complex the sentence is. The complexity of a sentence is directly proportional to the length of the sentence and the number of functional words in it. Therefore, there is a great difference in the frequency of terms used in these two texts, and its circulation characteristics show that the proportion of terms used in professional texts is higher than that in ordinary texts.

3. Translation Processing Algorithm of Complex Long Sentences Based on Multi-Strategy Analysis

3.1 Feature Extraction

Machine learning method is introduced into automatic estimation of translation quality in machine translation, and the selection of features has great influence on the system performance. However, the existing feature extraction only focuses on the feature mining of sentences, and the extracted features can not reflect the relative quality of terminology translation well. Because the rule base is manually built by many linguists, it is difficult to guarantee consistency. When the rule base reaches a certain scale, it is very difficult to further expand the rules. General decision dependency parsers are sensitive to long sentences. With the increase of sentence length, the parser performance drops sharply. Chinese verbs often use auxiliary words and adverbs to express time information, and the syntactic position of auxiliary words and adverbs is also an important factor that determines time information. Compared with the traditional phrase-based translation model, the hierarchical phrase translation model can deal with discontinuous phrases, and has certain generalization ability, and is not restricted by syntactic analysis.

In order to describe the expression degree of the term translation to the term information of the source language, that is, the fidelity of the term translation, we use formula (1) to approximately calculate the probability of translating English term \( f \) containing \( m \) words into Chinese term \( e \) containing \( n \) words in the absence of word alignment information.

\[
P(e_i|f_j) \approx \sum_{i=1}^{n} \log \left( \sum_{j=1}^{m} p(e_i|f_j) \right)
\]

(1)

In formula (1), \( p(e_i|f_j) \) is the translation probability between English terms and Chinese terms, which can be obtained by training word alignment information on English-Chinese term parallel corpus. At the same time, this method is also used to calculate the probability of generating terms in the source language, so as to further measure the degree to which the translated terms are faithful to their original texts.

In the regular matching method, the coverage rate of language rules is relatively low. And in the process of segmenting long sentences, only part-of-speech analysis and related information of parallel clauses are used, and the sentence pattern is only a simple part-of-speech combination of words, so its natural language features are one-sided. Moreover, the clauses after long sentence segmentation are still very long, and sentence nesting still exists [8]. In order to achieve better results in machine translation, this chapter improves the regular matching method. The algorithm flow chart of this chapter is shown in Figure 2.
Before extracting terms, an important task is to make statistics on corpus and get a statistical database, which is an important basis for extracting terms. It needs to go through the clause order adjustment module and the clause translation synthesis module to merge into a complete translation result and output it. Under the condition of considering the network voltage, through the logical connection between the ratio of the negative sequence current value and the positive sequence current value and the duration of exceeding the threshold value, the error is identified and distinguished from other network errors.

3.2 Long Sentences Are Concise

The main purpose of long sentence simplification is to merge some sentence components that affect segmentation. Legal association establishes the subordinate relationship between words; The subordination between words refers to the subordination of words in a sentence composed of direct connections. In each simplified rule, the left half of each rule is the precondition of the rule, which is expressed by regular expression. Sometimes, when describing the sequence of events, the causal relationship and the relationship between conditions and results, the order of narration in Chinese and English is just the opposite, so translation should be reversed. There are many models for machine translation of terms [9]. It only shows the contents of the corresponding fields, not real machine translation. This process is completed by manual addition. On this basis, this paper designs a terminology extraction program, which extracts the professional terms and then extracts the corresponding information segments from the improved statistical database. As shown in fig. 3.

Explain the meaning of terms and translating terms here are relatively simple, but the corresponding fields of extracted terms are displayed for the reference of those who read professional documents.

Based on the above knowledge representation, our simplified processing algorithm for segmentation of complex long sentences is to scan each component (word or punctuation mark) one by one for any input source sentence. In the quality estimation of machine translation, the author uses it to describe the fluency of terminology translation. However, the pure syntax-based SMT is limited by the inconsistency of bilingual syntactic structures, the over-generalization of terminators in generating rules.
and the large scale of generating rules, and the translation quality has not improved significantly. In the analysis strategy, the root searcher is constructed by machine learning method, and long sentences are separated by root information. Through such preprocessing, long sentences are transformed into short sentences, which reduces the complexity of sentences to a certain extent, and prevents the accumulation of errors to a certain extent. Extract sentences or phrases into formal expressions that can reflect their temporal characteristics. A translation template, which consists of real variables and variables inserted into each other in a certain order, and can be substituted by variables.

Let \( G = (V, E) \) be an undirected graph, \( V \) be a vertex set, and \( E \) be an edge set. \( Y = \{ Y_v \mid v \in V \} \), that is, each vertex in \( V \) corresponds to a random variable \( Y_v \), and its value range is the tag set \( \{ y \} \). If the observation sequence \( X \) is taken as the condition, each random variable \( Y_v \) satisfies the following Markov characteristics [10]:

\[
p(Y_v | X, Y_w, w \neq v) = p(Y_v | X, Y_w, w \sim v)
\]  

(2)

Where \( w \sim v \) denotes that two vertices are adjacent vertices in graph \( G \). Then \( (X, Y) \) is a conditional random field.

Given the observation sequence \( X \), the probability of a specific marker sequence \( Y \) is defined as [11]:

\[
\exp \left( \sum_j \lambda_j t_j(y_{j-1}, y_i, X, i) + \sum_k \mu_k s_k(y_i, X, i) \right)
\]  

(3)

Where \( t_j(y_{j-1}, y_i, X, i) \) is the transition function, which represents the transition probability of the labeled sequence \( Y \) at \( y_{j-1} \) and \( y_i \) under the condition of given observation sequence \( X \). \( s_k(y_i, X, i) \) is the state function, which represents the probability of \( Y \) being labeled \( y_i \) at \( i \) under the given observation sequence \( X \). \( \lambda_j \) and \( \mu_k \) are the weights of the above two functions, which are estimated from the training samples.

According to the grammatical features of Chinese, every sentence has only one root, and the root node plays an important role in a sentence. In the application of machine learning method, the information of root node can be used as an effective feature in dependency analysis. The accuracy of parameter estimation of translation model and language model depends directly on the size of corpus, and the translation effect ultimately depends on the coverage ability of probability model and corpus. Count the number of nouns, verbs and adjectives in the translation of terms, and then take the ratio of them to the number of all words in the translation of terms as the feature. The current input sentence "Source Sent" is approximately matched with the source text pattern of the qualified candidate pattern, and finally the corresponding translation is generated according to the solution pattern of the pattern. Ensure that readers accept the expression of the translated text psychologically. The coherence principle of Skopos theory requires language to be coherent, which requires the target language to conform to the grammar, syntax and expression habits of the target language, so as to avoid the influence of differences in language forms on the transmission and understanding of scientific and technological information.

3.3 Clause Segmentation

The previous steps mainly analyze and segment compound sentence effectively [12], but not all English long sentences contain parallel elements. And in long sentences with parallel clauses, the sentences after segmentation may still be very complicated. As mentioned above, dependency syntax can not only identify the syntactic component of each word in a sentence, but also reveal the dependency relationship between words, so it can express more information than phrase syntax. Each word is given
a number of knowledge such as concept category, HNC semantic symbol, sentence code, format code, expected knowledge, etc. From the concept category, it is convenient to indicate whether a word is a logical concept or a verb. If it is a verb, the sentence code can indicate which logical concepts the verb can be combined with.

In this paper, the score of the relative quality of each translation in the training stage depends on the human translation error rate (HTER) of the translation, which refers to the minimum editing times needed to manually edit the machine translation into the correct state (acceptable translation), including insertion, deletion, replacement and shift. With HTER as the benchmark for estimating the quality of terminology translation, automatic score estimation can intuitively reflect the workload of editing terminology translation into a basically correct state.

The definition of translation quality score in this paper is as follows

\[ \text{SCORE} = 1 - \frac{\text{HTER}(t,a_{-t})}{\text{LENGTH}(t)} \] (4)

Where, \( \text{HTER}(t,a_{-t}) \) represents the editing times of converting the system output translation into an acceptable translation, and the smaller the editing times of \( \text{HTER}(t,a_{-t}) \), the better the translation quality, and \( \text{LENGTH}(t) \) represents the length of the system output translation.

Split translation means that when some complex long sentences in German patents are quite different from those in Chinese in order of arrangement or expression, they are first divided into several small units according to sentence structure and translated one by one, and then rearranged according to Chinese habits. After analyzing the approximate matching between the input sentence Source Sent and the source text pattern of the candidate pattern and generating the corresponding translation, it is necessary to construct the complete translation of the current sentence according to the translation pattern of the candidate instance pattern. Usually, as long as the central word is found in the sentence, it is easy to understand the meaning of the sentence. Therefore, the head word is very important.

The editing distance between two strings \( X \) and \( Y \) is only an integer value, so we need to define a formula to calculate two levenshtein percentages. According to the formula of editing distance, define the following calculation formula:

\[ \text{Sim}(X,Y) = 1 - \frac{D(X,Y)}{\text{Max}[\text{Length}_X, \text{Length}_Y]} \] (5)

In this paper, a similarity threshold is set to obtain translation examples. If the similarity between the source language of the term to be optimized and the translation examples in the case library is greater than the set threshold, the translation examples are extracted.

The generalization degree of examples can be high or low, that is, special language components (such as named entities) in bilingual examples can be represented by class labels [13]; The root of a sentence is borne by the word "development", and the sentence is divided into two parts with the word "development" as the boundary. There is no cross between the two clauses in the left and right parts. For a sentence that cannot judge the time, if it is the first sentence, let the time of this sentence = the time of the next sentence; If the next sentence is still unable to judge the time, let the time of this sentence be equal to the time of the next x sentences. If it is not the first sentence, let the time of this sentence = the time of the previous sentence; If the previous sentence is still a sentence that cannot judge the time, let the time of this sentence = the time of the previous x sentence; If the last x sentences are sentences that cannot judge the time, let the time of this sentence be equal to the time of the next x sentences. After such a clause is simplified, the sentence pattern becomes very short, which is much easier for subsequent machine translation.

4. Experimental Results and Their Analysis

Similarities between sentences can be manifested in different aspects such as semantics, structure, target characteristics and individual characteristics. According to analogical reasoning, the optimal
matching should satisfy the aforementioned constraints at the same time. This paper selects the bilingual sentence-level corpus of "Chinese-English Comparative Example Sentence Set (500 Sentences)" and "Double Sentence Alignment Corpus (the first 500 sentences in 1 200 sentences)", the bilingual text-level corpus of "All over America (Lessons 20-25)" and "Bilingual News on Inter-net (12 articles in total)", respectively Artificial statistical experiment results (Table 1.).

Table 1 Basic Information of Test Corpus

| Corpus                                      | Total number of sentence pairs |
|---------------------------------------------|-------------------------------|
| Open test 1 Chinese-English comparative examples | 300                           |
| Open test 2 Two-sentence alignment corpus (the first 500 sentences in 1 200 sentences) | 500                           |
| Open test 3 Bilingual Text-level Corpus of Walking Through America (Lesson 20 ~ 25) | 450                           |
| Open test 4 Bilingual news (12 articles in total) | 129                           |

Generally speaking, the effect of machine translation mainly depends on F value, recall rate and accuracy rate, and the segmentation results are analyzed and predicted by these three criteria. The formula of recall rate is shown in figure 6:

\[
\text{Recall rate} = \frac{\text{Correct identification number}}{\text{Actual existing number}} \times 100\% 
\]  

(6)

The formula of accuracy rate is shown in figure 7:

\[
\text{Correct rate} = \frac{\text{Correct identification number}}{\text{Total number of identifications}} \times 100\% 
\]  

(7)

The formula for calculating the f value is shown in figure 8:

\[
F = \frac{2 \times \text{Recall rate} \times \text{Correct rate}}{\text{Recall rate} + \text{Correct rate}} \times 100\% 
\]  

(8)

Among them, in the above three formulas, the correct recognition number represents the correct number of sentences matched by regular regular expressions in the regular matching algorithm; The actual existence number is the number of components that exist in the text and should be detected; The total number of recognition refers to the total number of sentences that match the regular expression.

Table 2 lists some experimental results. The experimental results mainly consider recall rate, accuracy rate and F value.

Table 2 Experimental Results of Long Sentence Segmentation

| Method                      | Recall rate | Correct rate | F value |
|-----------------------------|-------------|--------------|---------|
| Long sentences are concise  | 92.3%       | 88.3%        | 88.6%   |
| Compound sentence segmentation | 90.6%      | 87.4%        | 89.2%   |
| Clause segmentation         | 92.7%       | 84.1%        | 84.1%   |
| Synthetic approach          | 92.1%       | 83.6%        | 85.2%   |

It can be seen from Table 2. that the recall rate of each step has achieved good results, and the application of comprehensive method has improved the accuracy of sentence segmentation. However, the accuracy rate of comprehensive method is not very high, so the method needs further improvement. According to the test of 10,000 online corpus, there are 25% super-long sentences, and the translation accuracy of these sentences is almost zero without simplified segmentation; Through the simplified segmentation of long sentences, the detection results of 650 punctuation-free super-long sentences and super-long complex sentences are shown in Table 3:
Table 3 Detection Result

| Number of complex long sentences | Simplify the number of sentences after segmentation | Translation comprehension rate | Segmentation time | Translation time        |
|---------------------------------|-----------------------------------------------|-------------------------------|------------------|------------------------|
| 650                             | 1672                                          | 52.33%                        | 5s               | 14 minutes and 12 seconds |

Because there are long and complex attributes behind the objects in the original sentences, it will be confusing and difficult to understand if we adopt the principle of cis-translation. Therefore, it is necessary to use the split translation method to translate the three parallel components first, and then discuss them separately. Although they are not directly comparable, it can be seen that the translation effect of the integrated system has reached or even exceeded the translation level of the best international news field.

5. Conclusion
With the progress and development of science and technology, machine translation methods are increasing. In the process of natural language processing, no matter what type of machine translation system is used, the more accurate, standardized and vivid the natural language expression, the higher the quality of machine translation will be. In this paper, aiming at the problem of long sentence translation in patent translation, the algorithm of sentence segmentation is used to deal with Chinese sentences. The long sentence is divided into one or more independent clauses, and a rule-based translation system is used to translate the clauses separately. After adjusting the sequence of clauses and synthesizing the clauses, the results of large sentence translation are finally given. The time information of Chinese sentences is analyzed and translated into English based on multi-strategy method. Experimental results show that this method can effectively solve the problem of temporal transformation in Chinese sentence time information analysis and Chinese-English machine translation.

The work of this paper is only an attempt in the translation of complex long sentences. The results are not very good, and some problems can not be completely solved. We hope it can be used as a reference for researchers and developers in related fields. In the future research, we can further expand the database to improve the recall rate, on the other hand, we can make further research on the processing of terms.

References
[1] Rao Yanyan. Research on Segmentation Translation Algorithm of Complex Long Sentences Based on Semantic Features. Journal of zhoukou normal university, vol. 37, no. 01, pp. 101-105, 2020.
[2] Zhang Xiaoxue. Hot spots, domain composition and prospect of English translation of The Analects——Visual analysis of co-words based on CNKI academic journals from 2001 to 2017. Shanghai Translation, vol. 00, 0no. 005, pp. 69-74, 2017.
[3] Chen binrui, Zhou Yuqin, yuan ying. a study on the translation of Jin Yong's works based on newmark's translation theory. Changjiang series, vol. 000, no. 006, pp. 37-39, 2020.
[4] Xu Ning. Study on the Application and Strategy of Comparative Translation in College English Translation Teaching. Journal of Huangshan University, vol. 020, no. 004, pp. 126-129, 2018.
[5] Shen Da. Subjective research on English translation of the derogatory construction "Haoge NP" — —Based on the comparative analysis of A Dream of Red Mansions translated by Yang and Huo. Journal of Shenyang University (Social Science Edition), vol. 20, no. 004, pp. 477-482, 2018.
[6] Zhang Rui. Research on English Long Sentence Translation Based on English Language Characteristics. Journal of hubei correspondence university, vol. 30, no. 001, pp. 162-163, 2017.
[7] Sha Jiu, Feng Chong, Zhang Tianfu, et al. Research on Tibetan-Chinese bi-directional neural machine translation with multi-strategy segmentation granularity. Journal of Xiamen University (Natural Science Edition), vol. 059, no. 002, pp. 213-219, 2020.
[8] Zhong Shuting. A Study on the Application of Grammatical Analysis in English Long Sentence
Translation. Contemporary Teaching and Research Essays, vol. 000, no. 006, pp. 118-118, 2018.
[9] Jiang Yue, Wang Letao. Research on Translation Teaching Practice Based on Domestication and Foreignization. Educational Theory and Practice, vol. 38, no. 33, pp. 51-53, 2018.
[10] Xue zhengmao. Analysis of translation strategies and methods in government work report. northern lights, vol. 000, no. 006, pp. 98-99, 2019.
[11] Chen dinggang. On the translation skills and strategies of English long sentences. English square, no. 5, pp. 11-13, 2018.
[12] Zhu Liyun, Xu Jingxian. A summary of CET-4 paragraph translation research——A visual analysis of co-words based on CNKI journals from 2013 to 2018. English Teachers, vol. 019, no. 016, pp. 13-16, 2019.
[13] Jin Fang. Analysis of translation strategies for dealing with cultural differences in English-Chinese translation. Think Tank Times, no. 07, pp. 188-189, 2020