Research Article

English Translation of Chinese Topic Sentences with Gap Subject Based on Internet Environment

Yanmei Geng

1School of Foreign Languages and Literature, Tianjin University, Tianjin 300350, China
2School of Education, Tianjin University, Tianjin 300350, China

Correspondence should be addressed to Yanmei Geng; lindagym@tju.edu.cn

Received 17 March 2022; Revised 4 April 2022; Accepted 6 April 2022; Published 9 May 2022

Abstract

Machine translation is showing an increasingly broad application prospect as the Internet becomes more widely used. The method based on the principle of maximum entropy is primarily used in this paper to determine the subject clause for translation. Its feature set does not necessitate extensive linguistic knowledge, and it is less reliant on it than other methods. It can also add arbitrary features and flexibly combine a large amount of scattered and fragmented knowledge. As a result, the maximum entropy model is used as the classifier in this paper, and lexical and sentence features are used to effectively combine the rules and statistical knowledge. The subsentence recognition process is broken down into three stages, with different features extracted at different stages and the maximum entropy model applied multiple times. At each stage, the classifiers are trained.

The translation accuracy and recall rate of Chinese gapped subject topic sentences are improved by more than 5% in a maximum entropy model, according to the results. This method of feature description is particularly useful for identifying sentence endings in Chinese subject topic sentences with gaps. Its translation is closer to the correct translation, indicating that the proposed method’s basic concept—the rationality of segmentation—is correct. The algorithm based on the maximum entropy model performs better in practice.

1. Introduction

Today’s society has evolved into an information society. With the rapid development of information technologies, such as computers, software, and networks, human beings have entered the information age. Against this background, people in different regions, cultures, and languages communicate more and more frequently, and people need to process information in various languages quickly and extract correct information from them, a useful part of itself [1]. As a method to explore the intelligent behavior of human beings to understand natural language, natural language understanding research has made great progress in the past two decades. The research content in the field of natural language processing is very rich, mainly including machine translation, information retrieval, and question answering system. With the popularity 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 [2].

Subject with vacancy Subject is a relatively general concept, and there is no uniform definition of what kind of sentence is the subject. There is no uniform standard for how long the length is and what kind of structure is a complex structure. Many researchers have proposed methods of subject segmentation [3] with gaps. However, different people may have different views on the rationality of the segmentation results. The ambiguity of related concepts in long sentences makes it difficult to study the subject of the gap.
Many existing methods can obtain certain results, but whether the segmentation results are correct and why they are correct, different people have different opinions, and the judgment of the results is more subjective. That is to say, there is a lack of standardized and objective measures to evaluate the results, which will also limit or affect the application of the method [4].

The principle of maximum entropy has been successfully used in many tasks in natural language processing and has achieved good results. The analysis and research of long English sentences are not all for English to Chinese English translation. For the processing of English long sentences in English-Chinese translation, there is a key factor to consider, that is, whether the processed result is conducive to the generation of Chinese translation [5]. This point can be used as an important reference for evaluating the segmentation results. Therefore, how to make the segmented English clauses easy to generate Chinese translation is another difficulty in current research. Both rational rule-based methods and empirical statistics-based methods need to select appropriate features for subject-gap classification. The choice of rules or statistical methods has a great relationship with the actual application [6]. At the same time, in long English sentences, there are both deep-level and shallow-level linguistic phenomena; both deductive and inductive methods are used. The number of selected features, the pros and cons, will greatly affect the effect of the method. However, there is no unified view on how to obtain features and which features to select, which is one of the reasons why long sentence analysis is challenging in the field of English translation.

The novelty of this paper is that, in order to correctly analyze various complex sentences in Chinese topic sentences with vacant subjects, the task of identifying subordinate clauses must first be completed. In natural language processing, whether the choice of feature templates and the appropriateness of feature representation [7–9] will have a significant impact on labeling, so this paper combines the characteristics of this model with the maximum entropy principle method and proposes the use of lexical features, sentence features, and other features. Experiments show that the method of feature description grammatical rules is extremely effective, particularly in the recognition of clause endings.

Section arrangement is as follows: Section 1 summarizes the principles of various translation machine models in the world; Section 2 builds a recognition model and algorithm of Chinese gapped main sentences based on the principle of maximum entropy; Section 3 uses sample analysis to conduct experimental analysis; and the fourth Section is the summary of the full text.

2. Related Work

The research and practice of translation informatization has a long history, and the most typical and well-known one is the study of machine translation. On the other hand, the application of information technology in translation work has greatly improved the labor productivity of translation and changed the labor production mode of translation to a considerable extent, thus becoming an important factor that plays a decisive role in the industrialization of translation.

A series of new achievements and new initiatives in machine translation research have sparked hope for a resurgence of the field since the mid-1970s. During this time, the Guo Xin-developed practical machine translation system TAUM-METTEO was officially implemented, the European Community Multilingual Machine Translation Program was proposed, and Japan proposed the Asian Multilingual Machine Translation Program ODA. [10]. People’s desire for computers to perform language translation has grown stronger in recent years, thanks to the rapid development and popularization of computer network technology. A noisy channel model was used in the original statistical translation method. The method assumes that sentences in one language will be deformed by a noisy channel, resulting in the presentation of sentences in another language at the other end of the channel. This method demonstrates that any sentence in any language could be a translation of a sentence in another language, but the likelihood varies [11]. Jiahui proposed a neural network-based translation method. The artificial neural network method, like the memory-based method, can map a source language sentence to a target language sentence, and its network model can be obtained through corpus training [12]. A method of instance-based translation was proposed by Wang. This method necessitates lexical, syntactic, and even semantic analysis of a known corpus, as well as the creation of a translation instance library. The sentence is preprocessed first, then the similarity is analyzed against the translation instances in the instance database, and finally, the translation of the translated sentence is obtained based on the translation of the similar instance [13]. Siqi was the first to propose machine translation, and 1 was the first to conduct a Russian-English machine translation experiment [14]. Teng proposed the famous SCHRLDU model [15] and published the theory of process semantics. Zhou et al. proposed the conceptual graph theory of language knowledge framework representation (ConceptualGraphs) and established the semantic network theory [16]. The FINL general syntax analysis system [17] was designed by Lu and includes three translation engines: a knowledge-based translation engine, an instance-based translation engine, and a word-translation-based translation engine. A tool is provided in the project to aid alignment, including the function of operating paragraphs and sentences in the article, specifically moving up paragraphs, moving down, splitting, merging; and moving up, moving down, splitting, merging sentences, among other things. Although the sentence similarity algorithm should be used for corpus alignment, due to the characteristics of the corpus of the Chinese translation company, that is, the corpus alignment is very neat, as a result, the sentence similarity method is not used, and only two articles with an alignment relationship are segmented, displayed after segmentation, and sent to the editor for alignment.
There are pros and cons to either method, and no method is absolutely superior to or even completely replaces all other translation methods. Ultimately, it is the true and correct research direction to achieve the integration of various translation methods.

3. Construction of Chinese Subject Recognition Technology with Gap

3.1. Problems in Subject and Related Term Identification.

Long sentences in unrestricted domain texts are not as thoroughly analyzed by today's syntactic analysis technology as simple sentences, and the accuracy rate is far from meeting people's needs. Syntactic analysis, on the other hand, is critical for many natural language processing tasks. To clearly analyze a sentence's structure, we must first understand its hierarchical structure, and clause recognition serves as a bridge to further analysis. The problem of chunk recognition, which falls under the shallow syntactic analysis category, is commonly referred to as the problem of labeling [18]. As a result, their research ideas resemble part-of-speech tagging. Although clause recognition falls into this category, the complexity of clauses means that this type of problem can no longer be compared to labeling problem research. This method realizes the segmentation of Chinese subject with gaps in three steps, as shown in Figure 1.

To begin, the effect of shortening the sentence is achieved by combining the sentence components that affect segmentation in the sentence. The main purpose of shortening is to combine two or more words into a single part of speech, reducing the number of “words” in a long sentence. Certain sentence components that affect segmentation, such as prepositional phrases and gerund phrases that can be used as adverbial components, quotation marks or brackets that can be used as appositional components, and punctuation-connected components, may be present in a given long input sentence. Combine simple phrase components in a new way. Incorporating these elements not only “shortens” sentences but also makes subsequent processing steps easier [19].

Second, locate and segment the sentence’s compound sentences. Long sentences use either punctuation or conjunctions to connect compound sentences. The parallel English clauses sometimes do not conform to the logic of Chinese sentences due to the logic habits of Chinese sentences. In English, for example, the result clause is frequently placed first in a sentence, followed by the cause clause. The order of the clauses after segmenting should be considered when segmenting compound sentences to reflect the Chinese language habit. This step is primarily based on the proper categorization of sentence types, so that all long sentences with coordinating sentences are placed in the appropriate categories. Furthermore, this step will segment the adverbial clauses with relatively independent grammatical structure, which are primarily found at the start or end of the sentence [20].

Finally, identifying the subordinate clauses present in the subsegmented clause in the second step makes the sentence more concise. These clauses can be subject clauses or attributive clauses. By identifying the existing clauses, the structure of the sentence can be simplified to a greater extent, which is more conducive to the processing of the machine translation system.

After conducting manual reading analysis on a large number of English and Chinese long sentence aligned corpora, we found that English and Chinese are very different in the use of long sentences. In Chinese, a long sentence is often composed of several shorter clauses. These short clauses are usually separated by punctuation marks to express a complete meaning together. Due to this feature, we can easily mine the information of these punctuation marks, segment sentences from certain punctuation marks, and hand them over to the machine translation engine for translation. But in English, it is different. English speakers tend to use a long whole sentence, and although this long sentence can contain many long phrases and modifying clauses, English writers often connect these elements directly before and after the main clause, rather than using punctuation marks and separate them from the main clause. This phenomenon is particularly evident in English official texts, such as news, scientific literature, and patent literature. The use of such long sentences brings great difficulties to sentence segmentation and simplification. In the existing methods, except using the rule method, it seems that it is difficult to effectively find the split points between these subordinate clauses and long phrases and the main clause.

3.2. Identification of the First Sentence of Chinese with Gaps Based on the Principle of Maximum Entropy. Clause start recognition plays an important role in the whole clause recognition system, because its prediction result directly determines the accuracy of the latter two stages, very low results. In this stage, the specific task is to predict whether the candidate word in each sentence is the first word of the clause [21]. In this paper, candidate words refer to the starting words of basic phrases in each sentence, and phrases composed of single words are also used as candidate words, as shown in Figure 2.

The features of each candidate word are added to the feature set in the training phase, resulting in the creation of a feature space and an event space. One of the events is a structure that contains the current word, all of its features, the clause's labeling category, and the number of times the event has occurred. There are two types of features in sentence start recognition: lexical features and sentence features. The lexical feature, for example, uses the sliding window method to produce three different types of features: part-of-speech tagging, phrase tagging, and part-of-speech sequence combination. Verb information, function word information, sentence structure, punctuation information, and other situations are all included in sentence features. The function word information primarily refers to whether the subordinate clause's leading word appears in the word, which requires special attention because when it precedes the attributive clause, there is frequently a preposition in front of it, and the beginning of the clause at this time is not the preposition [22]. The combined features of verb phrases and punctuation are referred to as sentence
Enter a sentence

Does it need to be streamlined?

Y

Short sentences

N

Does it contain a compound sentence?

Y

Parallel sentence segmentation

N

Does it contain a subject?

Y

Subject segmentation

N

Output sentence

Reduced rules

Compound sentence rules

Clause rule

Figure 1: Flowchart of subject segmentation based on regular matching.

Training corpus

Feature extraction

Main sentence start recognition

Main sentence end recognition

Multiple discrimination

Generate subject candidate set

Main sentence head recognition model

Main sentence ending recognition model

Multiple discriminant model

Generating subject candidate set model

Subject tagging

Figure 2: Detailed frame diagram of subject recognition system.
structure information at this stage. Other situations in the sentence features are the features that are supplemented by continuously analyzing the errors in the experimental results, as the maximum entropy model can add arbitrary features and flexibly combine many scattered and fragmented knowledge.

The principle of maximum entropy is to find the probability distribution with the largest entropy under the premise of known partial information, and the known partial information is embodied by feature constraints. Therefore, this paper introduces the concept of characteristic function and sometimes abbreviated as characteristic. Usually in the principle of maximum entropy, the feature is represented by a binary function, as shown in (1).

\[
f_j(a, b) = \begin{cases} 
1, & \text{if : } a = S \& h(b) = 1, 1 \leq j \leq k, \\
0, & \text{else.} 
\end{cases} \tag{1}
\]

Here, \(S\) represents the first word of a clause, and the function \(h(b)\) represents a predicate. The predicate is also a binary function, and the value is \([0, 1]\), which corresponds to whether a piece of useful information appears in the current context.

The mathematical expectation of the feature \(f_j(a, b)\) for the probability distribution \(p(a, b)\) determined by the model is

\[
E_p(f_j(a, b)) = \sum_{a,b} p(a, b)f_j(a, b). \tag{2}
\]

The mathematical expectation of the feature \(f_j(a, b)\) for the empirical rate distribution \(\hat{p}(a, b)\) determined by the model is

\[
E_{\hat{p}}(f_j(a, b)) = \sum_{a,b} \hat{p}(a, b)f_j(a, b). \tag{3}
\]

The feature constraint in maximum entropy requires these two expectations to be equal, and it becomes such a set of constraints for a set of features. However, in natural language processing, people often need to ask for the conditional probability \(p(a|b)\), and the conditional entropy should be used for the maximum entropy model, as shown in

\[
H(p(a|b)) = \sum_{a,b} p(b)p(a|b) \log p(a|b). \tag{4}
\]

The expectation of the feature \(f_j(a, b)\) for the model \(p(a|b)\) should be written as

\[
E_p[f_j(a, b)] = \sum_{a,b} p(b)p(a|b)f_j(a, b). \tag{5}
\]

Among them, \(E\) is the normalization factor, other parameters have the same meaning as the original model, and the number of features in the general natural language field is very large, so the estimation of model parameters is particularly important. The main task of the identification of complete clauses at this stage is to mark out the clauses in the insult sentence by combining the previous prediction results, including the case of multiple clauses [23]. Depending on the task characteristics, this task cannot simply be regarded as a labeling problem.

3.3. Multiple Discrimination Module. There are two parts to the multidiscrimination module. It determines whether the main sentence’s first word is a double sentence start, and then whether the double sentence first word is a triple case. Because there are few cases of triple or more in the corpus used, and only a small proportion (less than one tenth of the total number of words in the training corpus), the recognition in this paper only achieves triplex. Furthermore, using the maximum entropy model to determine the multiplicity can directly determine whether it is a single, double, or triple, that is, a multiclassification problem; however, experimental results show that this method has a poor effect in steps, so this paper uses a two-step method [24].

All the first words of main sentences in the training corpus are used as research objects, and the words originally marked as double or multiple cases in the sentence are regarded as double cases, followed by the feature training model. This module incorporates Chuan’s features and continues to follow the two major categories of features established in the previous two stages, but the sentence features have been supplemented with features that can better reflect dual information. It means that not only the sentence structure information for the entire sentence is included in the sentence structure feature but also the sentence structure information between the current word and each sentence ending word predicted after it is added. The number of occurrences of the prepredicted sentence start and end words in this sentence, as well as their relative positions to the current word, are added as features. Because the training and recognition processes are identical to the main sentence start recognition process, they will not be discussed further [25].

The research object was transferred to the first words of the double sentence that had been predicted in order to distinguish between double and triple, and the triple and above cases in the original correct subject labeling were considered triple. Others agree with the initial judgment’s double sentence. Because triple and above situations are uncommon in the training corpus, and because the model is inaccurate due to the limited training corpus, the experimental effect is not much better than double discrimination, and there are almost no triples in the test corpus, this part is an optional module. As a result of the aforementioned circumstance, the inclusion of this module in the system is primarily for system integrity.

This issue is also discussed in the aspect of subject recognition. The relevant smoothing methods that have been applied at present are mainly represented by the truncation method and the Gaussian prior method. Thus, setting the intercalary value causes features smaller than this value to be deleted. But later, this hypothesis has been proved, and these removed features may affect the accuracy of the entire
model. Therefore, most people use Gaussian prior method. Since the maximum entropy method can be understood as the maximum likelihood solution in essence, the Gaussian prior method is modified on this theoretical basis, as shown in

\[ E_p[f_i(a, b)] = E_p[f_i(a, b)] \times e^{\frac{\alpha_i + \delta_i}{\sigma^2}}. \]  

(6)

In this paper, in the first identification part of the main sentence, the influence of the smoothing algorithm on the model is observed by adjusting the size of the variance.

In the first step of the regular matching segmentation method, the sentence components that affect the segmentation are merged. The merged objects include the components that are connected by punctuation and should not be segmented at the punctuation, the prepositional phrase whose initial part acts as an adversarial component and the apposition, sentence components in parentheses and quotation marks of slang components, and phrases connected by adverbs and verbs, adjective phrases, noun phrases, etc. [26]. In the component merging step of this chapter, the abovementioned more complex components merged in the regular matching method are excluded. In this way, on the one hand, the impact of incorrect merging on segmentation can be reduced, and on the other hand, the complexity of regular expression formulation can be reduced. Incorporating only simple phrase components and absolute nonsegmentation components containing natural segmentation points at the same time can reduce the method’s dependence on sentence parts of speech, while retaining the efficacy of streamlining long sentences [27].

4. Experiments and Analysis

4.1. Experimental Setup. This paper randomly selects 500 Chinese subjects with gaps in a corpus as the test corpus to conduct segmentation experiments and verify the improvement of the translation quality by the segmentation results. Part of the experimental results of long sentence segmentation is evaluated by using the accuracy rate, recall rate, and F value. The corpus selected in this paper provides the correct Chinese translation. In order to evaluate the effectiveness and rationality of the method in this paper, this paper evaluates the quality of the translation of the source sentence without segmentation and the translation obtained by applying the algorithm in this paper. Higher translations are closer to the correct translation. Machine translation systems, especially rule-based translation systems, tend to be more prone to word-for-word or word-for-word translation when they cannot properly analyze the sentence to be translated. If the two languages are similar in order, word order, etc., the changes in the results of machine translation will be relatively small; otherwise, the quality of the translation will be greatly reduced. The identification of long sentences has reference significance for translation accuracy. Taking the number of consecutive words in the text as the starting point, the similarity and distance between the tested translation and the reference translation are calculated to evaluate the machine translation. If it is applied to the automatic scoring system of English-Chinese translation test, it can solve the problem that the traditional manual scoring test completely depends on the subjective judgment of the rater. Parts of speech can largely reflect the structure of sentences. The segmentation of long sentences requires an accurate grasp of the sentence structure. The algorithm in this paper combines regular expressions with the parts of speech of words in the sentence to describe the specific sentence structure and determine the position of the segmentation point. Table 1 lists common parts of speech and their meanings.

<table>
<thead>
<tr>
<th>Part of speech</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>Conjunction</td>
</tr>
<tr>
<td>CD</td>
<td>Quantifier</td>
</tr>
<tr>
<td>NN</td>
<td>Noun</td>
</tr>
<tr>
<td>IN</td>
<td>Preposition</td>
</tr>
<tr>
<td>JJ</td>
<td>Adjective</td>
</tr>
<tr>
<td>VB</td>
<td>Verb</td>
</tr>
<tr>
<td>RB</td>
<td>Adverb</td>
</tr>
<tr>
<td>WDT</td>
<td>Guide word</td>
</tr>
</tbody>
</table>

4.2. Segmentation of Long Sentences with Gapped Subjects in Chinese Based on Regular Matching. In the long sentence reduction step, most of the errors are caused by part-of-speech tagging errors. This is because the formulation of the rules depends on the pattern of the sentence, which consists of the parts of speech of the words in the sentence. If the part-of-speech tagging is wrong, the rules cannot match such sentence components, and the corresponding merging operation cannot be performed. For example, the part-of-speech tagger mistakenly marks a verb as a noun, and the method in this paper combines the verb into a certain component, then the logical structure of the sentence is likely to be damaged due to this combination, so the subsequent segmentation steps cannot match the result. The results are shown in Figure 3.

As can be seen from Figure 3, each step has achieved good results, especially the recall rate of each step exceeds 90%. The experimental results show that the long sentence segmentation method based on regular matching can effectively realize the segmentation of long sentences. However, there are some errors in each step of this method, including merging errors and segmentation errors.

In the long sentence reduction step, most of the errors are caused by part-of-speech tagging errors. This is because the formulation of the rules depends on the pattern of the sentence, which consists of the parts of speech of the words in the sentence. If the part-of-speech tagging is wrong, the rules cannot match such sentence components, and the corresponding merging operation cannot be performed. For example, the part-of-speech tagger mistakenly marks a verb as a noun, and the method in this paper combines the verb into a certain component, so the logical structure of the
sentence is likely to be damaged due to this combination and the subsequent segmentation steps cannot match the results.

Selecting different features of the same model will bring different effects. Therefore, this paper conducts experiments on the feature combination problem in the part of sentence start recognition of long sentences with gaps. Each type of feature is represented by its symbol, as shown in Figure 4.

The test results show that when identifying the head of the main sentence with a vacancy, the recognition accuracy does not improve with the number of feature templates used, and the combined effect of lexical features and verb information is slightly worse than lexical features alone. Adding function word information based on the combination outperforms using only lexical features, indicating that the
information provided by function word features is better suited to identifying the main sentence's head with gaps. Furthermore, the results of the experiments show that when the part-of-speech sequence combination feature is included, the recall improves. This demonstrates that this feature captures the essence of the model, particularly when comparing the last two lines of the boundary essence of the subject with a gap, which demonstrates that sentence structure information is not very important in the identification stage of the subject with a gap’s first sentence.

4.3. Error-Driven Chinese Subject Segmentation with Gaps. After the sentences processed in the component merging step are segmented and error corrected, the experimental results shown in Figure 5 are obtained. The evaluation of segmentation results in this paper is based on the concept of segmentation rationality.

As can be seen from Figure 5, the component merging experiments yielded higher accuracy, recall, and F value. In this step, fixed collocation phrases, compound conjunctions, compound leading words, and common simple phrases are combined. These combined phrases cannot span prepositions, verbs, and gerunds. Rules that span natural segmentation points can only deal with common simple phrases. As much as possible, the segmentation will not be affected by the merging of sentence components.

It can be seen from the Figure 6 that both the training samples and the test samples only need to consider the phrase boundary words, which can achieve better results. Comparing the four cases, respectively, it will be found that when only the phrase boundary words are considered as the research object in the training model, if the test sample contains every word, the effect will be very poor. The reason for the analysis should be that the training model cannot extract the features of all words. Information is especially nonphrase boundary words, so errors may occur when discriminating these words.

4.4. Error Correction Experiment. Error correction is performed on the extracted long sentences. The main purpose is to remove useless noise from the corpus. The main steps include removing the news source information at the head of the sentence, removing special symbols, removing repeated sentences, etc. and using the method proposed in this paper to mark the comma position. The experimental results are shown in Figure 7.

It can be seen from Figure 8 that the size of the sample set will directly affect the recognition rate when other conditions remain unchanged. Therefore, it can be known that the size of the sample set extracted is not as large as possible, and the change of scale and the recognition rate are not linear, but there may be a peak, and the recognition rate can reach a maximum value under the specific sample set size.

The error-driven algorithm selects natural cut points as the basis in the component merging step. All sentence components that contain natural segmentation points but are clearly unsegmentable are merged. Because the number and types of natural cut points are many, they are limited. By fully examining the linguistic phenomena in actual sentences, the merging of components can be implemented more comprehensively. In the algorithm based on regular matching, more components are merged because it needs
to provide services for the subsequent steps, and the merging rules are more complex than those based on the error-driven algorithm. This makes it difficult to guarantee that the combined components will not contain split points that should be split. Once the merging is wrong, the result of segmentation will destroy the logical structure of the sentence, which will affect the translation result. In the subsequent segmentation steps, the error-driven algorithm takes into account many linguistic features related to segmentation. On the one hand, it ensures the number of segmentations, and on the
other hand, it provides linguistic support for error correction. In the algorithm based on regular matching, if the matching fails, the segmentation will not be performed.

4.5. Influence of Translation Processing of Chinese Gap-Subject Topic Sentences. Although many natural language processing tasks need to solve the performance degradation caused by sentence length and complexity, due to different tasks and different fields of text corpus, there has been no accurate definition of Chinese topic sentences with gaps. In addition, in order to study the segmentation and processing of long sentences, the first problem to be
solved is to clarify the research object, that is, to quantitatively analyze long sentences and determine the specific standards of long sentences. To this end, we set up a controlled experiment to analyze the effect of translated sentence length, as shown in Figure 9.

This paper argues that this impact characteristic is related to our strategy for constructing the dataset. When constructing the test set, we did not consider the sentence length but only the number of commas, so there are many sentences of shorter length in the test set. Even if the sentence length is long, because the condition of the number of commas is set, it is difficult for these sentences to have the possibility of adding new commas, so the number of commas is not added. The third is that after adding the segmentation module based on rule recognition, the accuracy and recall rate are slightly improved, which proves that the method based on the topic sentence with gaps in Chinese has a certain supplementary effect.

5. Conclusions

Based on the maximum entropy as the basic classifier, this paper hopes to further improve the system performance through ensemble theory. First of all, this paper proposes a subject recognition feature for the maximum entropy model by learning the traditional subject recognition technology 1. The feature can be divided into two categories: lexical features and sentence features. Sentence features include sentence structure, function word information, verbs and head and tail information, punctuation information, and other aspects of five aspects. Among the lexical features, the part-of-speech sequence combination feature can well express the subject boundary information, which draws on the idea of feature fusion. The key point of this method is to obtain a different new training set according to the sampling with replacement of ideas, then use the maximum descendant model as the basic classifier again, and fuse the results of each classifier to obtain the final prediction result. In this paper, the error correction of the final result is carried out, and the effect of this feature is particularly obvious in the stage of sentence ending recognition. The sample size of the number of classifiers is not as large as possible, and the change of the number scale and the recognition rate are not linear, but there may be a peak, that is, under the condition of a certain number of classifiers, the model performance is the best. The experimental results show that in a maximum entropy model, the translation accuracy and recall rate of Chinese vacant subject topic sentences are improved by more than 5%. Finally, the experimental results of complete subject tagging also prove that the algorithm-based subject recognition method can achieve a higher recognition rate.

It is difficult to accurately describe the real text in the limited training set, so the trained model always deviates from the real text. The subject recognition problem is a bridge between lexical analysis and deep syntactic analysis. Its recognition accuracy not only affects the effect of the subsequent analysis but also is directly affected by the front-end results. Therefore, according to the relevant research in this paper, it is believed that an appropriate algorithm is selected to extract complete and accurate. For the extracted features, the method of feature fusion can be considered, that is, before the feature is generated, the original feature template is combined to generate a new composite template.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author does not have any possible conflicts of interest.

Acknowledgments

This study was supported by the Ancient Chinese Characters (English Version) within the Chinese Academic Translation Project category.

References


Functional Equivalence Theory,” *Campus English*, vol. 21, article 2, 2018.


[Tengbo, “Translation skills for images in Chinese-English poetry from the perception of functional equivalence,” *Shanghai: Education Frontier*, vol. 11, article 1, 2017.]


