Hindawi
Mobile Information Systems
Volume 2022, Article ID 3090115, 10 pages
https://doi.org/10.1155/2022/3090115

Research Article

English Long Sentence Segmentation and Translation Optimization of Professional Literature Based on Hierarchical Network of Concepts

Di Shao and Rui Ma

College of Foreign Languages, North China Institute of Aerospace Engineering, Langfang, Hebei 065000, China

Correspondence should be addressed to Di Shao; shaodi6666@126.com

Received 7 June 2022; Revised 4 July 2022; Accepted 17 July 2022; Published 25 August 2022

Academic Editor: Jiafu Su

The structure of English long sentences is generally complicated, with complicated logical levels, many parallel elements, many modifiers and conjunctions, and long post attributives. Moreover, pronouns in English long sentences often need to be judged according to the context. Complex English long sentences are very common in English and even run through the whole English article. The analysis results of these complex long sentences will seriously affect the quality and readability of machine translation. In this paper, HNC (hierarchical network of concepts) method is improved to realize the segmentation of long sentences, so as to simplify sentence patterns. According to the characteristics of professional literature, this paper puts forward a translation optimization method of professional literature combining HNC with statistics, which greatly improves the recognition efficiency of unknown words by extracting professional terms. The results show that the word segmentation method based on HNC and statistics proposed in this paper has achieved a good word segmentation effect in the open test environment, with an accuracy rate of 93.38% and a recall rate of 94.51%. The conclusion shows that our method can make full use of the knowledge of the source tree database, thereby improving the accuracy of the syntactic model on the target tree database.

1. Introduction

In recent years, the domestic English translation business has been developing constantly, from the translation of English works in the past to the diversified English translation of English movies, TV dramas, news, etc. As a whole, great progress has been made. Machine translation has become an effective way to solve English translation. However, due to the complexity of language knowledge and the limited cognition of language rules, there is a big deviation between the accuracy of machine translation and language artistic conception [1, 2]. At present, the machine translation algorithms of English complex long sentences are based on syntactic analysis, multistrategy analysis, and corpus translation, which mainly focus on word meaning checking and semantic feature processing [3, 4], but these machine translation algorithms have low translation accuracy, high recovery rate, and poor reliability of translation results. Therefore, this paper studies the segmentation of English long sentences based on HNC (hierarchical network of concepts) and the optimization of professional literature translation, so as to realize the automatic proofreading of English translation and improve the efficiency and quality of English translation.

Professional literature translation involves multidisciplinary fields. Theoretically, it is appropriate that professional literature should be translated by professionals. However, the reality is that most professional people may not be proficient in foreign languages or they know foreign languages but lack basic translation skills, so it is difficult for them to be competent translators. In English translation, how to solve the problem of long sentence translation is a key point worthy of attention. From the perspective of English language structure, a long sentence means that the

components in a sentence are complex; for example, it may consist of several clauses or cover many other modifiers. Duan et al. put forward the concept of document processing automation based on word frequency statistics. It establishes a theoretical basis for selecting important words and abstract sentences [5]. Giannella et al. provide information of different levels of knowledge according to users' personal interests and try to shield this information that users do not like. HNC reasoning becomes the key technology of knowledge organization and knowledge query in intelligent information retrieval [6]. Matsuuchi et al. modeled the problem as finding the dependency tree with the highest probability (score) from a directed multiple graph (complete dependency graph) [7]. Chen et al. proposed a dependency analysis method based on integer linear programming [8]. They used the first-order graph model and added some strong constraints inspired by linguistics. Their model needs the constraint of exponential number, because every possible ring needs to introduce a constraint. Corpus from different fields plays a decisive role in statistical models. Therefore, even if a system has excellent effect and precision of word segmentation in its own domain, it often fails to achieve satisfactory results for word segmentation in other domains.

Complex English long sentences are very common in EST, some of which are even as long as dozens of lines, including hundreds of words, and contain many clauses and nonpredicate verbs. These clauses and phrases are interdependent and have very distinct semantic hierarchical relationships. In HNC, the phrase corresponds to the concept of semantic block. As the next semantic unit of a sentence, semantic block is the function of sentence type, and it plays a very important role in sentence type. The effective segmentation of long sentences can be used in the actual machine translation system, which can simplify the sentence structure and improve the overall performance of the machine translation system. In addition, it can also be used in the fields of natural language processing such as information retrieval and text classification. At present, under the premise that there is little research on word segmentation technology of professional literature and the word segmentation technology of professional literature is immature, the research on translation optimization of professional literature has great practical significance.

2. Related Work

2.1. Research on Machine Translation Technology. With the development of machine learning methods [9], artificial intelligence, computer technology, and linguistic theory, people have a deeper understanding of the background, objectives, and application prospects of machine translation, so machine translation has made great progress again.

Luo et al. believe that the biggest barrier to machine translation is semantic ambiguity. The diversity of language semantics requires that machine translation must have a set of grammar and syntax systems that can fully analyze language structure and semantic problems [10]. However, the conversion rules and syntactic analysis models applied at that time obviously could not solve all kinds of complicated language problems. Motoyuki and Kazuhiro discussed traditional machine translation methods based on dictionaries and conversion rules and example machine translation methods based on parallel corpora [11]. Condon et al. believe that any language problems faced by translation can be attributed to word meaning problems at the lexical level and structure problems at the grammatical level [12]. Wen et al. put forward a diagonal alignment model and an alignment model based on hidden Markov model. The diagonal alignment model can better solve the influence of small-scale corpus on translation results, and the alignment model based on hidden Markov model can make alignment smoother in general [13].

Wei et al. put forward a vocabulary domain labeling method based on the annotation information in dictionaries, which uses the annotation information of words in general dictionaries to label the domain of words, expanding the scale of existing HNC [14]. Wu et al. put forward the principle of machine translation method based on semantic unit theory, which regards the translation between natural languages as the conversion between different expressions of the same semantics in two natural languages. Firstly, semantic analysis is carried out in the source language to obtain sentence meaning expressions, which are then substituted into semantic units of the target language to generate sentences in the target language [15]. Liu et al. have studied the classification system of machine translation specialty, the mapping of specialty dictionaries to specialty classification systems, and the mapping of International Standard Classification ICS standards to specialty classification systems [16].

2.2. Professional Literature Translation Research. Over the years, many scholars at home and abroad have done a lot of research work in the field of term extraction. The methods used are rule-based methods, statistics-based methods, and the combination of statistics and rules. Jiang et al. introduced the concept of mutual information to measure the combination ability of two words and extracted various combinations of words [17]; Wang et al. adopted the method based on pure statistics to automatically extract terms from open corpus and jointly applied two statistical parameters, mutual information and log-likelihood ratio, to the automatic term extraction algorithm, extracting the seeds of two-word words to be expanded from corpus [18]. Geng adopts word-based maximum entropy model and word-based conditional random field model to automatically extract terms and transforms the problem of term extraction into the problem of term labeling, thus realizing an automatic term extraction system based on the combination of word-based conditional random field model and rules. The accuracy of the system is 89.9%, and the f value is 88.4% [19].

Kedar proposed a global training method using linear model for transition-based model, and the training algorithms of graph-based and transition-based methods tend to be consistent [20]. Tubau found through detailed comparison that the error distribution of the two methods is different. Compared with the graph-based method, the

transition-based method has high accuracy on the dependency arc with short distance (the distance between the core word and the dependency word) and on the dependency arc with far distance from the root node (from this arc to the root node in reverse) [21]. Pereira proposed a method to reorganize the results of multiple models [22]. The main idea is to establish a new dependency graph according to the analysis results of several models. The weight of each arc in the dependency graph is obtained by voting multiple model results. Ge proposed a fine-grained fusion method [23]. The idea of this method is to fuse the features used in transition-based and graph-based models and train the weights together.

3. Research Method

3.1. Segmentation of English Long Sentences by HNC Algorithm. Due to the extensive use of relative words, conjunctions, prepositions, etc., English long sentences usually contain a large number of coordinate elements, and the relationship between the main clause and the clause, as well as the clause and clause, makes the sentences complicated and difficult to grasp. Therefore, when translating English long sentences, translators often need to adopt the split translation method; that is, the middle participle, phrase, and even clause of the original text are split from the original text and then translated to facilitate the overall arrangement of the sentence. English is a language that emphasizes hypotaxis. It is customary to use conjunctions to show the relationship between sentences and the components in sentences by explicit means, while Chinese pays attention to parataxis, which mainly depends on the logical connection between sentences and components in sentences. No matter what kind of English long sentence it is, these methods and steps are involved in the translator's use of split translation.

Compared with short English sentences, long English sentences involve many subclauses, and the relationships among these subclauses are complicated. However, if we ignore the correlation, it will directly lead to readers' deviation in understanding the whole sentence. Actually, in order to express the meaning more clearly, people will make more use of the mixture of complex sentences and compound sentence. Such sentences are called mixed long sentences. In this sentence structure, sentences will contain more parallel structures, and these parallel structures will also have their own subordinate structures. Therefore, English long sentences should have the following characteristics: long sentences, many words, many sentence components, and many main or clause components. Moreover, the sentence structures are complex and changeable, and some of them have nested structures, which contain more grammatical and semantic information.

The HNC statistical algorithm used in this paper is mainly used to improve the feature selection algorithm, and the core metric of feature selection is the feature weight. If a concept is repeatedly mentioned in the article, it is often closely related to the main object of the article, and it is also of great value for category judgment. Judge whether the analysis results meet the design requirements of symbol system, how much they meet, and the next improvement direction. In addition, the consistency of each knowledge base is tested quantitatively. In this way, the process of machine translation will involve the ordering of English sentences. There are also adverbial clauses with independent grammatical structure, which are mainly located at the beginning or end of a sentence.

In HNC method, this paper mainly analyzes sentences in three steps, and the process is shown in Figure 1.

The specific process is as follows:

(1) The part of speech of sentences is marked, and then the parts of speech are merged according to the regular expression matching rules, so as to achieve the purpose of simplifying sentence components. In this way, the number of "words" in the whole sentence is reduced, and then merging these components will bring convenience to the subsequent processing of sentences.

(2) Analyze compound sentence. In English sentences, a common problem is the word order before and after compound sentence; that is, after English is translated into Chinese, the translation must conform to the logical order of Chinese. For example, in common adverbial clauses of cause, the result in an English sentence is usually at the beginning of the sentence, while the reason is at the end of the sentence.

(3) If there are clauses in the segmented sentence, further segmentation is needed. These clauses can be subject clauses, object clauses or attributive clauses, etc.

Through the above three steps, try to make the sentences more concise and shorter, which will bring more convenience to the subsequent machine translation work.

The higher the level of a concept is, the more abstract it is and the wider the scope it covers. The concept covers too wide a range, and its significance to classification is relatively small. Use a generalization formula to measure:

$$F_R(c) = 1 - \frac{\max_{i=1}^{n}\left(F_t\left(S_i\right)\right)}{\sum_{i=1}^{n} F_t\left(S_i\right)}, \tag{1}$$

where $F_t(S_i)$ represents the number of times that the subconcept itself and all its subordinate subconcepts appear in the article and n represents the total number of subconcepts of this concept c.

In order to be like an artificial topic analyst, the text topic can be determined only according to several concepts that have strong predictive effects on the topic, and the hierarchical order of the text topic can be ensured. Semantic network is used to represent the hierarchical concepts of topics, and the concepts that have strong decisive effect on topics are selected as the central nodes in conceptual reasoning network.

Information gain is a criterion that is often used in machine learning to measure the importance of a certain predictor. First of all, the evaluation function is used to get a

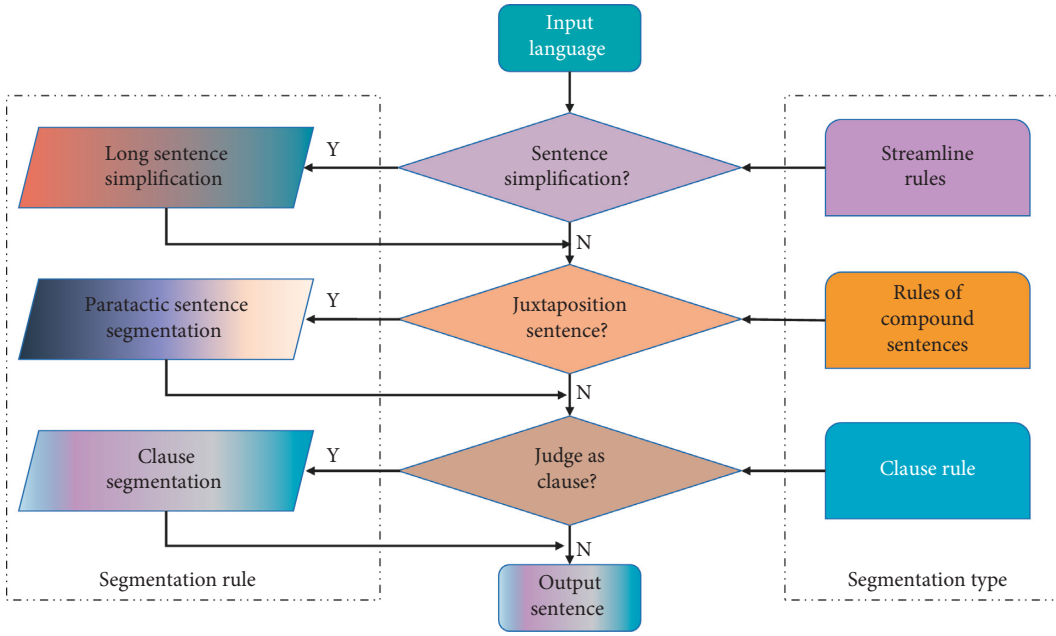

Figure 1: Streamline process based on HNC long sentences.

score for all the predictive words, and then they are sorted according to their size. Finally, according to the size of the threshold, it is decided to retain the predictive words that better reflect the characteristics of the theme. The evaluation function is as follows:

$$\begin{aligned}\mathrm{InfGrain}(F) &= P(F)\sum_i p(w_i|F)\log\frac{p(w_i|F)}{p(w_i)} \\ &+ P(\overline{F})\sum_i p(\phi_i|\overline{F})\log\frac{p(\phi_i|\overline{F})}{p(\phi_i)}.\end{aligned} \tag{2}$$

Here, F is a word, $P(F)$ indicates the probability of the occurrence of the word F, $\overline{F}$ indicates that the word F does not appear, $p(\phi_i)$ indicates the probability of the occurrence of class i, $p(\phi_i|\overline{F})$ indicates the probability of the occurrence of the word F, and ϕ_i indicates the probability of its occurrence.

$\mathrm{InfGrain}(F) > w$ words are the core words, and InfGrain is defined as the subordinate degree of the subject words. The number of core words is limited by obtaining a larger w.

The algorithm is based on the semantic knowledge base of HNC. Each word in the knowledge base gives the knowledge of HNC concept type, cluster code, etc. The type of the word is judged from the concept category, and the cluster code is the logical combination unit of the word matching.

Cosine similarity model is an important model commonly used to measure the word meaning difference between two phrases. It is based on multidimensional space, and the difference between two vectors is represented by the cosine of the angle between two vectors. If the cosine between two phrases is larger, the smaller the angle between two semantic vectors is, the closer the meaning of two phrases is. On the contrary, if the cosine value between two phrases is smaller, the semantic difference between the two phrases will be greater.

Let two phrases in the same corpus be multidimensional semantic vectors u, v, and let $u = [a_1, a_2, \ldots, a_n]$, $v = [b_1, b_2, \ldots, b_n]$, n be the dimensions of the vectors; then the English translation similarity $\mathrm{Sim}(u, v)$ between the two phrases is calculated by the following formula:

$$\begin{aligned}\mathrm{Sim}(u, v) &= \frac{u \cdot v}{\|u\| \times \|v\|}, \\ &= \frac{\sum_{i=1}^{n} a_i \times b_i}{\sqrt{\sum_{i=1}^{n} a_i^2 \times \sum_{i=1}^{n} b_i^2}}.\end{aligned} \tag{3}$$

In order to ensure the efficiency of global search algorithm based on dynamic programming, the graph-based model needs to make strong independent assumptions. It is assumed that, in the dependency tree corresponding to a sentence, only the dependency arcs existing in some special structures are interrelated and influence each other, while other dependency arcs are independent of each other. Under this assumption, the score of a dependent tree is decomposed into the sum of the scores of some subtrees.

$$\begin{aligned}\mathrm{Score}(x, \mathrm{d}) &= w \cdot f(x, \mathrm{d}), \\ &= \sum_{p \subseteq \mathrm{d}} \mathrm{Score}_{\mathrm{subtree}}(x, p),\end{aligned} \tag{4}$$

where p represents a subtree allowed by an independent hypothesis. p contains one or more dependent arcs in d. $\mathrm{Score}_{\mathrm{subtree}}(x, p)$ represents the score contributed by p. $f(x, \mathrm{d})$ is the aggregated syntactic feature vector corresponding to (x, d), and w is the weight vector of syntactic features.

Through explicit means such as morphological changes, relative words, and conjunctions, the relations between words, sentences, and clauses in the language are closely connected, so that the sentence structure is complete and is standardized, thus forming the syntactic features of English hypotaxis, as well as the features of English long sentences with complex structure, many logical levels, many modifiers, and many parallel components.

3.2. Optimization of Professional Literature Translation. Professional words can be divided into two categories: common words and technical terms. Compared with the general corpus, the boundaries between these two types of words are clearer. Among them, common words appear frequently, which reflects the writing norms and sentence characteristics of patent corpus, but their vocabulary is not large. First of all, a term is a fixed or semifixed word or phrase that is closely combined. It should be a linguistically established word. However, it is different from common words. The main difference is that it is a language unit with strong domain characteristics and is used in a specific professional field. Therefore, terms have three characteristics: tight combination, language completeness, and domain.

The part of speech words can also be determined by disambiguation according to the context information, but there is no new word that has never appeared in the training corpus, and naturally there is no context information that can be associated with it. Therefore, the processing of new words in part of speech tagging faces greater challenges. The training data corresponding to each affix is sparse and the part of speech of Chinese affixes has great ambiguity. Because Chinese words are shorter than English, some existing metagrammatical features may not be applicable to Chinese. Statistical methods are used to predict the part of speech of unknown words, so there is no need to add artificial customized grammatical rules. Therefore, when the training corpus is large enough, statistical methods can be applied to any language and have strong adaptability and practicability.

HNC divides cognitive structure into local and global associations. Local association refers to the association at the lexical level, which can be simply summarized as follows: dividing concepts into abstract concepts and concrete concepts, expressing abstract concepts with quintuple and semantic network, and expressing concrete concepts with anchored expansion approximation.

Specifically, it is to translate English long sentences directly according to the fixed order, without considering too much linguistic and artistic effects, mainly to ensure that the translation is clear and accurate or even to change jumbled long sentences into multiple short sentences in the translation process. According to the limited objects of auxiliary semantic blocks, the semantic blocks and even sentence types are further analyzed to reduce the difficulty of semantic block analysis and anatomy.

Technical terms are the linguistic expressions of concepts in various disciplines, the achievements of scientific research, and the crystallization of knowledge language in the course of human progress. In order to clearly explain the functions and technical characteristics of patents, it is necessary to use a large number of technical terms, and even some terms are self-coined. The domain difference between professional literature and general corpus determines that the word segmentation method in general domain is not suitable for professional literature. According to the inherent characteristics of professional literature, this paper adopts the word segmentation method based on HNC and statistics. The extraction process of technical terms in this paper is shown in Figure 2.

There are two steps to extract professional terms:

(1) Firstly, according to the rules of word formation, the rules of term extraction are summarized, and the candidate terms are evaluated and selected by using algorithms and forbidden word lists, so as to obtain preliminary professional terms.

(2) Secondly, aiming at the problem that low-frequency technical terms are difficult to identify, the preliminary technical terms are used as the template training text, and the conditional random field model is used to construct the term extraction template, so as to finally extract meaningful low-frequency technical terms and improve the extraction accuracy of technical terms.

The basic idea is to inspect one training example at a time. According to the current weight vector $w^{(k)}$, the prediction result of the current instance is obtained, and, according to the error in the prediction result, the current feature vector is updated to obtain a new feature vector $w^{(k+1)}$. The algorithm determines the final weight vector through multiple iterations. Every iteration needs to traverse all training examples.

According to the current weight vector $w^{(k)}$, the approximate optimal dependency tree of top $-K$ is obtained by using the decoding algorithm based on column search and stored in $\widehat{\mathrm{d}}^{K}$. The update criteria are

$$\begin{aligned} &\min\left\|w^{(k+1)}-w^{(k)}\right\|,\\ &s.t.\forall \mathrm{d}\in \widehat{\mathrm{d}}^{K} w^{(k+1)}\cdot f\left(\left(x^{(j)},\mathrm{d}^{(j)}\right)-f\left(x^{(j)},\mathrm{d}\right)\right)\geq \mathrm{loss}\left(\mathrm{d}^{(j)},\mathrm{d}\right). \end{aligned} \tag{5}$$

That is, the current feature vector $w^{(k)}$ is adjusted minimally, so that, for any dependency tree d in top $-K$, the difference between the score of the correct dependency tree $\mathrm{d}^{(j)}$ and the score of d is not less than the error number of d.

When the encountered new words have both prefixes and suffixes, how to select more effective affix information to improve the accuracy of new word labeling becomes the key. You can make a trade-off in the following ways:

(1) Choose the longer of prefix and suffix.

(2) Select the one with less parts of speech in the prefix and suffix distribution. If the prefix has a possible part of speech and the suffix has a possible part of speech, the prefix distribution is selected to estimate the lexical probability.

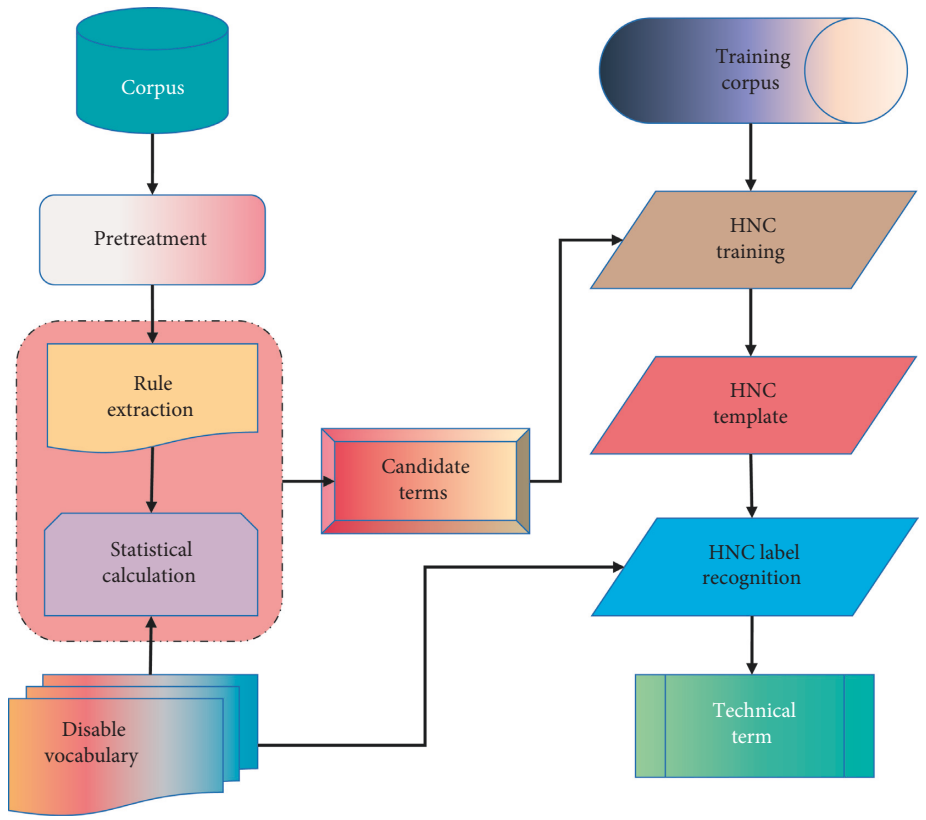

Figure 2: Technical term extraction process.

It is more useful to calculate the entropy of prefix and suffix distribution and then determine which one is more useful. The method of calculating entropy is used in part of speech tagging to measure how much information is necessary for independent data. For example, the formula for calculating the entropy of affixes is

$$S_a = -\sum_i \frac{n_{ai}}{N_i} \log_2 \left(\frac{n_{ai}}{N_i} \right), \tag{6}$$

where n_{ai} is the number of times that part of speech i appears in all words with a as affixes and N_i is the total number of times that affix a appears. Through calculation, the affix distribution with the smallest gun is selected to estimate $P(w_i|t_i)$.

By transforming the problem of part of speech tagging of new words into the probability of launching survival words, the unification of statistical methods can be well realized, and the problems of acquiring rule knowledge and building dictionaries can be avoided. Assume that the transition probability is only related to the previous state; that is, the part of speech c_j of x_j is determined by the part of speech of w_{j-1}. Namely,

$$P(c_j|x_j) \approx \sum_{m=1}^{M} P(c_m|w_{j-1}) P(c_j|c_m), \tag{7}$$

where M represents the total number of parts of speech categories. According to the Bayesian formula, the probability of lexical emission is

$$P(x_j|c_j) = \frac{P(x_j)}{P(c_j)} P(c_j|x_j). \tag{8}$$

In the formula, the number of times in which marked symbol c_j appears in the training expectation is represented by $C(c_j)$. The number of times that the symbol string $c_m c_j$ appears together in the training corpus is indicated by $C(c_m c_j)$.

4. Result Analysis

In some cases, the translation of English long sentences is limited by time, which will increase the difficulty of translation. For example, for simultaneous interpretation and consecutive interpretation in English, it is necessary for the translator to respond in time and translate the information during or after speaking. Accordingly, Chinese expression pays more attention to the overall semantic expression in language, ensuring harmony, conforming to artistic conception, and reflecting the characteristics of parataxis. However, English expression pays attention to the structure and form of sentences in language, showing the characteristics of hypotaxis. For the translation of long English sentences, it is necessary to translate English sentences that focus on hypotaxis from parataxis level.

In order to investigate the role of HNC statistical model in automatic text classification system, the system uses the same training text and test text to conduct classification experiments on HNC statistical model and morphology-

based model, respectively. The difference between the two systems is limited to feature weighting and selection. Take different numbers of training corpus and classify the same test set. Both HNC-based and word-form-based methods are used for comparative experiments. The experimental results are shown in Figure 3.

It can be seen that the classification accuracy of the system increases with the increase of training corpus, but, with the continuous increase of training corpus, the growth of the classification accuracy of the system slows down. No matter how much the training corpus is, the method based on HNC has higher accuracy than the method based on morphology. The highest difference is about 10% percentage points.

By the time of 80 articles, the accuracy rate of HNC-based method reached 92.35%, while that of morphology-based method was only 84.55%, with a difference of 7.8 percentage points. This shows that the classification method based on HNC has obvious advantages over the method based on morphology.

Table 1 is the comparison table of segmentation accuracy, recovery rate, and average cross-connection number between traditional algorithm and this algorithm. The accuracy rate and recovery rate of segmentation, respectively, indicate the accuracy rate of dividing complex English long sentences and the usage rate in specific translation, which are the basis for ensuring the accuracy of English long sentence translation and also important indicators. For the translation of complex English sentences, the higher the segmentation accuracy, the lower the average recovery rate, and the higher the translation accuracy.

It can be seen that the number of cross-connections of this algorithm is 17.46, which is 12.68 less than that of the traditional translation algorithm, indicating that the performance of this translation algorithm is better. Compared with traditional algorithms, this translation algorithm has higher translation accuracy and recovery rate, so it has higher practicability.

Figure 4 lists some experimental results. The experimental results mainly consider the recall rate, accuracy rate, and F value.

It can be seen that the recall rate of each step has achieved good results and the application of comprehensive method has improved the correct rate of sentence segmentation. However, the correct rate of comprehensive method is not very high, so the method needs further improvement.

English pays attention to the continuity of logical thinking and is used to connect the components of sentences by explicit means, so that the relationships among the components of sentences can be revealed, while Chinese pays more attention to parataxis. There will be some mistakes in word segmentation and labeling on the training corpus of this patent, and there will be some mistakes in the corresponding term recognition results. It is easy to divide the whole professional term or part of the term into a string with other words, and this phenomenon of word adhesion often occurs in the front and back boundaries of terms. Particularly, for long sentences, the analysis time of high-order models increases rapidly. This

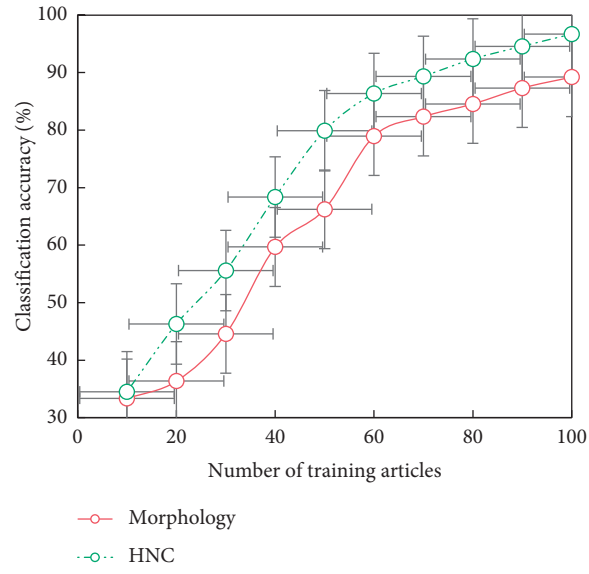

Figure 3: Classification result.

Table 1: Comparison results of various indexes between traditional algorithm and this algorithm.

	Traditional algorithm	Algorithm in this paper
Number of clauses after segmentation	1369	1683
Correct cut fraction	1224	1579
Cross-connection number	30.14	17.46
Average accuracy	89.41%	93.82%
Average recovery rate	7.16%	15.38%

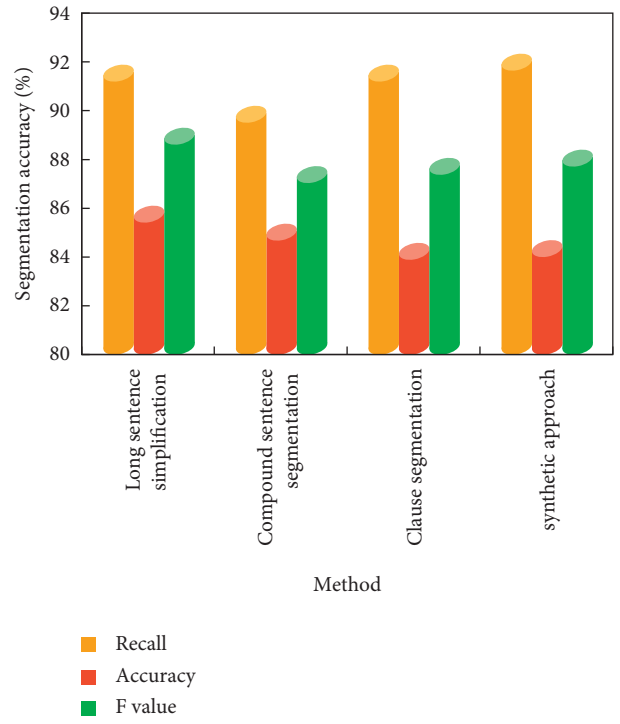

Figure 4: Experimental results of long sentence segmentation.

restricts the application of high-order dependency parsing model in high-level natural language processing systems, such as machine translation and information extraction. Therefore, according to the characteristics of Chinese, we propose a fast dependency parsing method based on punctuation.

The translation results are evaluated by the international general evaluation indicators BLEU and mteval-v13a. Table 2 shows the evaluation results of the three systems.

Under a standard answer, the BLEU values of this system and [10] translation system are 0.193, and 0.142, respectively. The BLEU value of the former is much higher than that of the latter. It can be seen that the translation quality of this system is obviously better than that of the comparative translation system.

The corpus used in this experiment is the Chinese patent data provided by the patent MT task of NTCIR-9 conference. In this paper, HNC and CRF++ tools are used to train the training corpus, so as to extract patent terms. The sampling statistical results of HNC and CRF algorithm terminology recognition are shown in Figure 5.

It can be seen that although the training accuracy of CRF model is lower than that of HNC, more technical terms are identified. In this experiment, 1863 preliminary terms were identified by HNC, and 2014 terms were identified by CRF model training. Compared with HNC, the vocabulary difference (number of changes) was 56893.

In this paper, the patent HNC is constructed by rule extraction and statistical learning, and there are 440, 247 entries in the constructed HNC. Figure 6 is the experimental result of segmenting all words in the test corpus in the open test environment.

It can be seen that, through word segmentation experiments on professional literatures in many fields, the general word segmentation tool in [17] has only 70.66% accuracy rate for patents and the effect is not satisfactory. The word segmentation method based on HNC and statistics proposed in this paper has achieved good results in the open test environment, with 93.38% accuracy and 94.51% recall in the open test.

The imperfection of the corpus makes it impossible to contain all linguistic phenomena, which leads to the deviation of the acquired linguistic information and the inability of correctly reflecting the real linguistic phenomena. It also causes serious data sparseness of the calculated model parameters, which affects the accuracy of labeling. Because the training corpus cannot contain all possible words and parts of speech, the language is constantly developing, and new words will be added. So, when there are a large number of new words, because the relevant information cannot be obtained from the training corpus, the probability of wrong tagging will be greatly increased, which will seriously affect the accuracy of tagging. Therefore, a better solution to the problem of new words is also a big problem that puzzles part of speech tagging.

For this reason, we extract the dependency structures corresponding to all segments from the training corpus. Figure 7 compares the efficiency of different methods in different sentence lengths.

Table 2: Evaluation of translation evaluation results.

System type	BLEU-4
Ref. [10]	0.142
Ref. [13]	0.176
This paper system	0.193

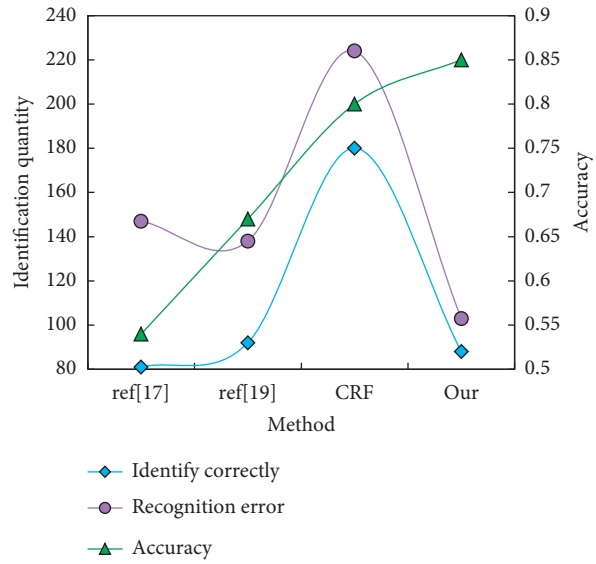

Figure 5: Sample results of term extraction experiment.

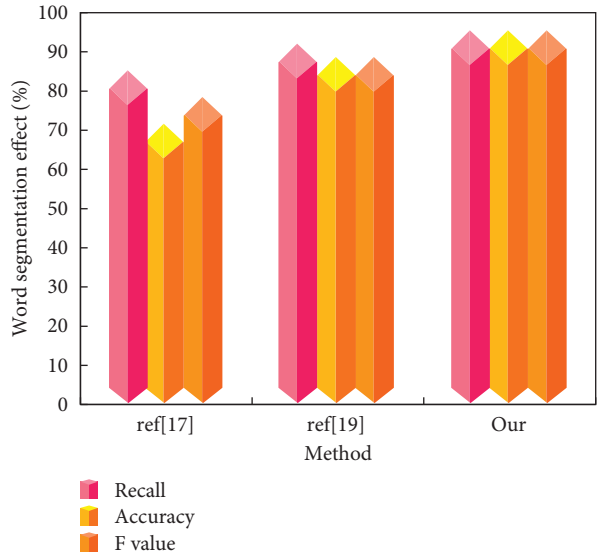

Figure 6: Experimental results under open test.

For the sake of clarity, we have omitted sentences longer than 110. Time analysis includes all operations from the input of sentences to the output of analysis results, such as feature extraction and decoding. It can be seen that the two-stage method greatly improves the efficiency of analyzing long sentences and the influence of sentence length on the analysis time is very small.

Because the average sentence length of the corpus is 25, we used three intervals: sentences less than 29, sentences

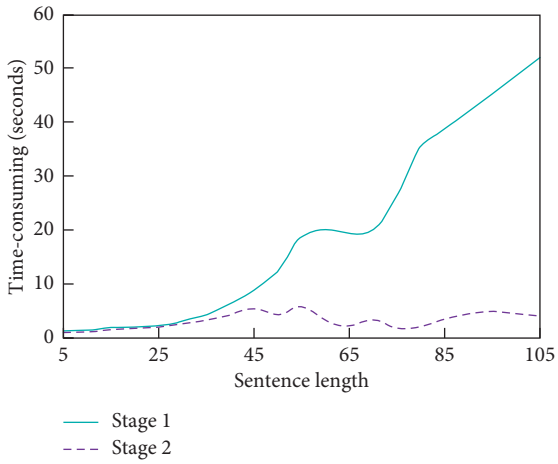

Figure 7: Efficiency comparison of different sentence lengths.

greater than 50, and sentences in the middle. It can be seen that, compared with the one-stage method, the two-stage method increases the syntactic accuracy with the increase of sentence length. For sentences longer than 28, the accuracy of the two-stage method is greatly improved.

Merging the converted source tree database with the target tree database to form a larger training data, finally, the syntactic analysis model is trained on the merged tree database. As the training data increases, the accuracy of the syntactic analysis model will naturally increase. However, it is a very difficult problem to deal with the inconsistency of annotation structures in different tree repositories. Through a clever strategy to reduce the noise contained in the transformed tree library, the performance of syntactic analysis can be improved. In the experimental part, we indirectly compare our method with theirs, and the results show that our method is more effective.

5. Conclusion

There are many long sentences in English, which are easy to be translated into blunt and dull Chinese due to the different syntactic features between English and Chinese, and sentence translation is the cornerstone of text translation, which makes English long sentence translation a major focus and difficult in translation. This paper applies HNC algorithm to English long sentence segmentation. A regular expression is used to describe or match a series of strings that conform to a certain syntactic rule; that is, it provides a way to match strings. The word segmentation method combining HNC with statistics can effectively improve the recognition rate of unknown words by extracting technical terms. The word segmentation method based on HNC and statistics proposed in this paper has achieved good results in the open test environment, with 93.38% accuracy and 94.51% recall in the open test. Careful error analysis shows that the joint model can better resolve syntactic sensitive part of speech ambiguities and correct resolution of these ambiguities can further help syntactic analysis.

Data Availability

The labeled dataset used to support the findings of this study is available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This research was supported by "Model Course of Moral Education-College English," the Education Department of Hebei Province (KCSZ202104S).

References

[1] K. Shaalan, A. Hendam, and A. Rafea, "Rapid development and deployment of bi-directional expert systems using machine translation technology," *Expert Systems with Applications*, vol. 39, no. 1, pp. 1375–1380, 2012.

[2] Y. Zhang, "Research on English machine translation system based on the internet," *International Journal of Speech Technology*, vol. 20, no. 4, pp. 1017–1022, 2017.

[3] K. N. Dew, A. M. Turner, Y. K. Choi, A. Bosold, and K Kirchhoff, "Development of machine translation technology for assisting health communication: a systematic review," *Journal of Biomedical Informatics*, vol. 85, pp. 56–67, 2018.

[4] H. Wu, "Multimedia interaction-based computer-aided translation technology in applied English teaching," *Mobile Information Systems*, vol. 2021, no. 5, pp. 1–10, 2021.

[5] G. Duan, H. Yang, and K. Qin, "Improving neural machine translation model with deep encoding information[J]," *Cognitive Computation*, vol. 13, no. 3, pp. 1–9, 2021.

[6] C. R. Giannella, R. K. Winder, and J. P. Jubinski, "Annotation projection for temporal information extraction," *Natural Language Engineering*, vol. 25, no. 3, pp. 385–403, 2019.

[7] K. Matsuuchi, T. Fukami, N. Naoe, R. Hanaoka, S. Takata, and T Miyamoto, "Performance prediction of a hybrid-excitation synchronous machine with axially arranged excitation poles and permanent-magnet poles," *Electrical Engineering in Japan*, vol. 150, no. 2, pp. 43–49, 2005.

[8] W. Chen, D. Kawahara, K. Uchimoto, Y. Zhang, and H Isahara, "Using short dependency relations from auto-parsed data for Chinese dependency parsing," *ACM Transactions on Asian Language Information Processing*, vol. 8, no. 3, pp. 1–20, 2009.

[9] X Chun and M.-Y. Zhang, "Multiple linear regression for extracting phrase translation pairs[J]," *Journal of Computers*, vol. 6, no. 5, pp. 905–912, 2011.

[10] Y. Luo and Y. Xiang, "Application of Data Mining Methods in Internet of Things Technology for the Translation Systems in Traditional Ethnic Books[J]," *Data Mining for Internet of Things*, vol. 8, pp. 93398–93407, 2020.

[11] H. Motoyuki and H. Kazuhiro, "Media processing technology for achieving hospitality while on the go.[J]," *NTT Technical Review*, vol. 13, no. 4, pp. 1–7, 2015.

[12] S. Condon, M. Arehart, D. Parvaz, G. Sanders, C. Doran, and J. Aberdeen, "Evaluation of 2-way Iraqi Arabic–English speech translation systems using automated metrics," *Machine Translation*, vol. 26, no. 1-2, pp. 159–176, 2012.

[13] X. Wen, "Hierarchical phrase machine translation decoding method based on tree-to-string model enhancement[J]," *Acta*

Technica CSAV (Ceskoslovensk Akademie Ved), vol. 62, no. 1, pp. 531–539, 2017.

[14] Y. Wei and Y. Guo, "Research and analysis of improved extraction based on information processing technology[J]," *Journal of Digital Information Management*, vol. 10, no. 2, pp. 142–146, 2012.

[15] X. Wu, Y. Xia, J. Zhu, L. Wu, S. Xie, and T Qin, "A study of BERT for context-aware neural machine translation," *Machine Learning*, vol. 111, no. 3, pp. 917–935, 2022.

[16] X. Liu, W. Wang, W. Liang, and Y Li, "Speed up the training of neural machine translation," *Neural Processing Letters*, vol. 51, no. 1, pp. 231–249, 2020.

[17] Y. Jiang and H. Liu, "Research on the construction of parallel corpus for the specific field of machine translation[J]," *Boletin Tecnico/Technical Bulletin*, vol. 55, no. 19, pp. 77–82, 2017.

[18] D. Wang, J. Su, and H. Yu, "Feature extraction and analysis of natural language processing for deep learning English language," *IEEE Access*, vol. 8, pp. 46335–46345, 2020.

[19] B. Geng, "Text segmentation for patent claim simplification via bidirectional long-short term memory and conditional random field," *Computational Intelligence*, vol. 38, no. 1, pp. 205–215, 2022.

[20] Y. Kedar, "Interaction between two determiner systems: the acquisition of English articles by a Hebrew-speaking child," *First Language*, vol. 39, no. 1, pp. 111–136, 2019.

[21] S. Tubau, "The asymmetric behavior of English negative quantifiers in negative sentences," *Journal of Linguistics*, vol. 56, no. 4, pp. 775–806, 2020.

[22] R. C. Pereira, "Epenthesis and deletion as strategies to acquire complex syllabic structures: strategies in the Interlanguage of Brazilian learners of German as a foreign language," *System*, vol. 98, no. 2, Article ID 102479, 2021.

[23] L. Ge, "Study on the translation techniques in English Chinese translation of passive sentences in educated[J]," *Linguistics*, vol. 3, no. 1, pp. 407–420, 2021.