Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

A Computational Approach to the Classification of Chinese Syntactic Structures Using a Large-Scale Corpus

Song Wang[1,] ✉

[1] *School of Chinese Language and Literature, Xinyang College, Xinyang, Henan, 464000, China*

ABSTRACT

Syntactic analysis is a basic work in the field of natural language processing, which explores the syntactic structures and their interaction relations in sentences. This paper first describes the basic approach of syntactic analysis, and explores the computational method of Chinese syntactic structure classification from large-scale corpus construction. Then, a grid-based large-scale corpus construction and distribution model is constructed. And the word embedding model BERT is used as the pre-trained language model, and the captured semantic features are input into the Bi-LSTM model to extract the contextual bidirectional sequence information, and the results of Chinese syntactic structure classification are obtained by the Conditional Random Field (CRF) processing. Through manual proofreading as well as the calculation of confidence level, the average correct rate of syntactic structure classification of the final Chinese canonical corpus is increased from 94.21% to 99.06%, which is an improvement of 4.85%. The syntactic structure classification accuracy of the BERT-Bi-LSTM-CRF1 and BERT-Bi-LSTM-CRF2 models with "complement structure" and "object structure" were higher than those of the BERT model, the Bi-LSTM-CRF model and the BERT-Bi-LSTM-CRF3 model with all syntactic structures. Meanwhile, the accuracy of the syntactic structure annotation method of BERT-Bi-LSTM-CRF model + manual differs from that of manual annotation by only 0.66%, and the average time spent is reduced by 37.04%, which reduces the workload of the annotators and improves the efficiency of the annotation, which verifies the validity and practicability of this paper's model in automatic classification of Chinese syntactic structures.

Keywords: large-scale corpus, BERT, Bi-LSTM, CRF, syntactic structure classification

1. Introduction

Syntactic structure refers to an organic framework formed by the interconnection and interaction of the components of a sentence in the syntactic plane, which contains five basic categorical forms: subject-predicate, paratactic, complementary, verb-object, and union [1]. These five forms can be

✉ Corresponding author.
E-mail address: andywong1985@126.com (S. Wang).
Received 05 August 2024; Revised 25 October 2024; Accepted 15 February 2025; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-164

understood as basic grammatical relations in the Chinese form. Syntactic structure is the core building block of the Chinese grammatical system, which is where the characteristics of Chinese grammar lie. Chinese sentences are directly composed of syntactic structure, except for the associative words in one-word sentences and compound sentences as well as interjections [2-4]. However, with the advent of the information age, the changes in people's life can be described as radical, new information constantly appearing around people, triggering people's new ideas, and then triggering the new syntactic structure of the language, prompting the rapid development of modern Chinese, and a large number of mutated syntactic structure of modern Chinese is spreading rapidly with the help of the Internet. In social life, young people have the strongest ability to accept new concepts, new things and innovations. Variant syntax is an emerging form of language in the new period, which crosses the conventional syntax and greatly satisfies the young people's mentality of seeking for differences and new things, and they can get a new feeling and experience from it [5-6]. And the expressive power and infectious power of variant syntactic structures are often stronger, most young people will frequently use new syntactic structures in order to pursue new feelings and experiences, and these variant formats are cohesive with special meanings in special contexts and have strong expressive power, and their large number of uses and popularity are related to the creativity of young people as the main language users [7-8]. In addition, avoiding complexity and seeking simplicity is a common psychology of human beings, and the process of language development can be said to be a process of striving to describe the changes of the objective world more precisely and efficiently, and this psychology of language users brings about the complexity of the classification of Chinese syntactic structure [9]. In addition, due to the international development of Chinese language, more and more countries teach Chinese as a second language, in the context of developed network language and multilingual learning, the in-depth analysis of Chinese syntactic structure in teaching is necessary, and learning to categorize the types of Chinese syntactic structure separately is a necessary way to master the correct understanding and use of Chinese language [10-12]. Therefore, in the context of the great changes in people's language structure and ideological cognition, a large number of new syntactic structures have emerged, and these changes are not only the result of social development, but also affected by the subject of language use. In order to better understand and use the Chinese language, it is an important topic to analyze the classification of Chinese syntactic structures.

In this paper, we investigate the computational methods for classifying Chinese syntactic structures using large-scale corpora. First, a grid-based large-scale corpus construction and distribution model is designed, and a credibility calculation method for the corpus is proposed. Then, based on the Tsinghua Chinese Treebank and the constructed Chinese canonical corpus, the BERT-Bi-LSTM-CRF model is utilized to classify Chinese syntactic structures. In order to verify the effectiveness of the model, the classification effect of this paper's model is compared with the BERT model and Bi-LSTM-CRF model. Meanwhile, the comparison with manual annotation proves that combining this paper's model with manual annotation can improve the efficiency of manual annotation while ensuring the accuracy of Chinese syntactic structure category annotation.

2. Syntactic analysis and classification of the syntactic structure of the Chinese language

2.1. Syntactic analysis

Syntactic analysis is a core technique in computational linguistics that solves grammatical problems by exploring grammatical associations between words as well as phrase constructions in sentences. This approach is crucial in the field of natural language processing. Syntactic analysis methods are mainly categorized into rule-based syntactic analysis, corpus-based syntactic analysis, and syntactic analysis combining rules and statistics.

2.1.1. Rule-based syntactic analysis. The rule-based approach is a knowledge-based rationalist approach. The approach is based on linguistic theory and emphasizes the linguist's knowledge of linguistic phenomena. There are mainly the following rule-based methods:

1) Method of adding syntactic markers. Syntactic analysis with added syntactic markers consists of a sequence of state transducers, and the transducers consist of regular formulas, i.e., the syntactic rules are in the form of finite state grammars. Most rule systems use this approach.

2) Automatic learning of syntactic rules. In the rule-based approach, the main difficulty lies in the acquisition of grammar rules there and the prioritization ordering between grammar rules. Therefore, scholars have proposed a conversion-based error-driven learning method, which obtains an ordered set of rules for recognizing basic noun phrases through learning. Another method for automatic acquisition of syntactic rules is to use the instance-based method or memory-based method in machine learning.

Rule-based syntactic analysis has the following drawbacks:

1) The rules are summarized from a closed corpus and are not ideal for open corpus.

2) The robustness of the analyzer is low, as evidenced by the fact that it often fails when meeting unexpected situations (e.g., raw words).

3) Difficult to represent knowledge of small granularity, and therefore not good at handling ambiguity.

4) Difficulty in ensuring consistency of rules and compatibility between linguistic rules.

2.1.2. Corpus-based syntactic analysis. It is because rule-based approaches have some shortcomings in some aspects that people are turning to large-scale corpora in an attempt to obtain less granular linguistic knowledge from them to support natural language processing systems for large-scale real texts.

In many cases, there will be more than one syntactic tree for a sentence, and we try to rank the results of these syntactic analyses statistically. We use “S” to denote the sentence to be analyzed, “T” to denote all possible syntactic trees, and G to denote a set of syntactic rules. The core of statistical (corpus)-based syntactic analysis is to define a probabilistic model "M", which is used to calculate the probability that sentence S corresponds to the syntactic tree T. The task of statistical-based syntactic analysis is to find the tree with the highest probability value, as shown in the formula:

$$P\left(t^*\right) = \arg\max P(t \mid S, G) \qquad Where \qquad \sum_{t \in T} P(t \mid S, G) = 1 \tag{1}$$

There are two approaches that are often used together in statistically based syntactic analysis: rule-driven and data-driven. In the rule-driven approach, a generated grammar is used to describe the language and kind of language being analyzed. When performing an ambiguous syntactic analysis, we use statistical information to rank the results of multiple post-selected syntactic analyses. Data-driven approaches can derive grammatical knowledge from a manually labeled training dataset (the Pennsylvania Treebank). When doing a syntactic analysis, our goal is to select the syntactic analysis with the most reasonable structure based on the training data so that it can be applied to the context of the sentence.

2.1.3. Syntactic analysis combining rules and statistics. In view of the increasing maturity of rule-based approaches and corpus statistical methods, scholars have proposed to utilize the respective advantages of rule-based and statistical methods to improve the ability of Chinese syntactic analysis.

The existing literature proposes to establish a flexible analysis system that integrates the advantages of the two. The rule-based approach is first utilized to establish the dependency network of sentences.

Then the dependency network is simplified by statistical and rule-based methods. Finally, the tree in the dependency network is evaluated and output.

Rules are mainly used for merging and pruning neighboring phrases. Statistical methods are mainly used for automatic labeling of lexical properties, automatic labeling of dependencies and evaluation of syntactic trees. A salient feature of this model is that the analysis of dependency grammars is regarded as a problem equivalent to lexical labeling, and a statistical model is used to accomplish both kinds of labeling.

2.2. Classification of Chinese Syntactic Structures

The syntactic structures of Chinese are mainly categorized as follows:

1) Para-corrective structure. A structure consisting of two parts, the first modifying word and the center word, and the relationship between the two is that of modifying and being modified. One part modifies or restricts the second part of the structure, such as “white dove”. “The word “white” modifies and restricts the word “dove”, which is the modifier, and the word “dove” is the main (center) word. There are mainly definite center structure, gerund center structure, temporal structure, quantitative structure and so on.

2) Stated Object Structure. e.g. Doing the laundry. The first part of the sentence is a predicate (indicating action or behavior), and the second part is an object (something connected with the first part). Commonly, there are verb-object structures such as: “to eat” (VP→VP NP). Double-object constructions such as: “Call me Xiaoduan” (VP→VP NP NP).

3) Complement structure. For example: I finished eating. Unlike "eating apples", in "eating apples", "apples" are the objects of "eating". And in "finished eating", "finished" is the result of "eating". The first part is the statement, and the second part is the complement.

4) Subject-verb construction. E.g. I know. The subject-predicate structure consists of two parts: the subject is the object of the statement, and the predicate is the statement of the subject, i.e., how or what the subject is.

5) Joint Structure. For example: Beijing, Shanghai, Harbin. It consists of several components of equal status in juxtaposition. Unlike the above sentence structures, the joint structure may consist of more than one part, not just two.

6) Conjunctive Structure. E.g. Run out and open the door. A conjunctive predicate structure is a format in which predicates or predicate structures are used together.

In fact, a sentence in Chinese may contain multiple structures, such as "there are many foreign students in Xiao Ming School" itself is a subject-verb structure, the subject "Xiao Ming School" is a positive structure, the predicate is a descriptive structure, and the predicate "foreign students" is a partial structure. The most commonly used sentence structure in Chinese is the subject-verb structure, that is, the subject-verb-object structure, and there are some manifestations in the relative clause, genital complement, and noun modifier structure. Numbers cannot directly modify nouns, and quantifiers need to be added.

3. A large-scale corpus-based model for categorizing Chinese syntactic structures

In order to utilize a large-scale corpus to classify Chinese syntactic structures, this paper applies a grid-based large-scale corpus construction and distribution model to construct a large-scale Chinese

corpus, and proposes a BERT-Bi-LSTMCRF syntactic structure automatic classification model applicable to Chinese corpus.

3.1. A grid-based model for large-scale corpus construction and distribution

In this paper, a grid-based large-scale corpus construction and distribution model is designed as shown in Fig. 1. Using this model, we can not only collect Chinese corpus from multiple ways and parallelize the processing, but also provide a release platform for corpus information based on the grid, so that the large-scale Chinese corpus constructed in this paper can provide corpus services for syntactic structure classification in a service way.

The Grid middleware in Fig. 1 is the core of Grid computing, which is responsible for providing remote process management, resource allocation, storage access, login and authentication, security and quality of service. In this paper, Globus Tookit 4 is chosen as the grid middleware.

The text information processing service platform, on the other hand, provides the basic functions of corpus collection and distribution, and it mainly provides specific services based on the content of the corpus in the corpus. In addition, in this computation- and service-oriented Grid, the nodes can be distributed everywhere and can physically take on multiple functions. The task scheduling of the nodes is dynamic and is done by a specific grid scheduling program. The Scheduling Manager in the Word Information Processing Service Platform provides this functionality, which controls the operation of the entire grid and is primarily responsible for the management of jobs.

Portal (Portal) is the interface and portal of an information grid to release information for the Internet. It can provide corpus information services and other text information processing services.

Wiki based on Information Grid is a platform for collecting and processing corpus information. On the one hand, it can collect raw corpus information, and on the other hand, it can provide cooked corpus checking for the other two corpus collection methods.

Users can be categorized into three groups according to their operating environments: grid nodes, users of non-grid nodes, and users of other handheld devices. Among them, computers that are grid nodes can both act as a node of the information grid to provide computational power to the grid, as well as utilize the grid environment built locally to directly access the data of the grid and utilize the grid for computation. Therefore, when users use this type of node, they can directly utilize the service interface to use the corpus services provided by the grid. Non-Grid nodes, on the other hand, need to interact with the Grid through the Portal, utilize the interfaces provided by the Portal to provide corpus for the Grid, and make general use of the service interfaces provided by the Portal to use the corpus services provided by the Grid. Users of handheld devices, on the other hand, can act as users of the Grid's corpus services and application services, and they mainly use the corpus and application services provided by the Grid. They can also build corpora like the other two types of users, but their computational power is limited and they can generally only be used as an input terminal.

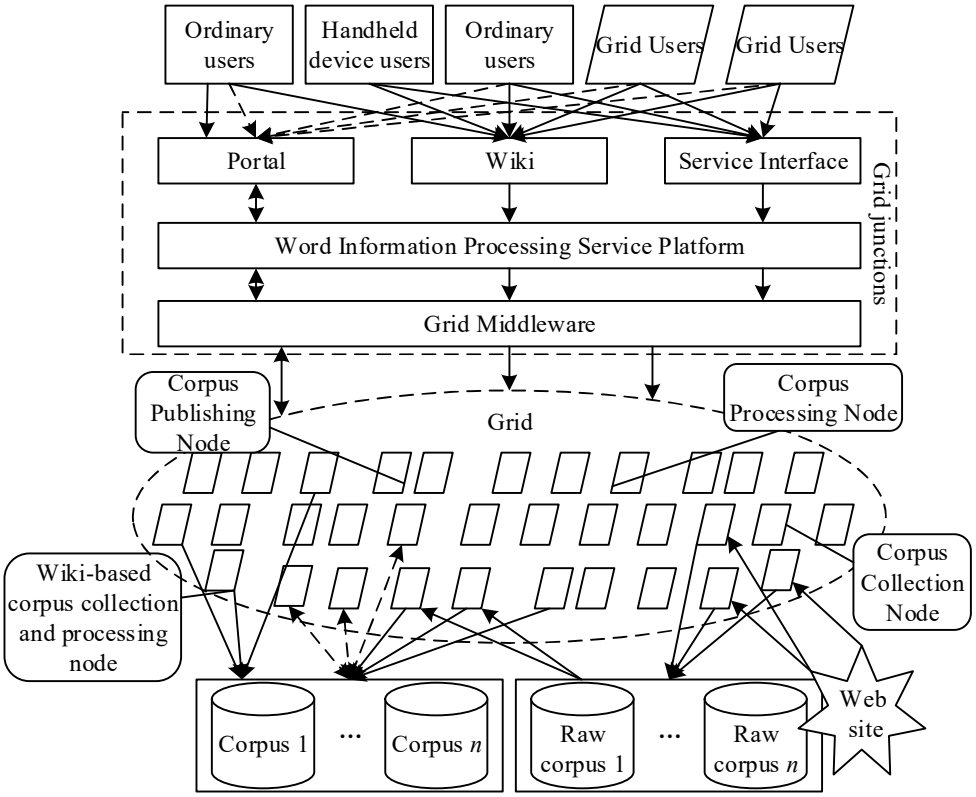

Fig. 1. Grid-based corpus construction and publication model

3.2. Credibility Calculation Methods for Chinese Corpus

The definition of credibility in this paper is as follows:

Definition 1: Credibility indicates the degree of trust that a thing or phenomenon is true, and is expressed in terms of T, with values ranging from 0 to 1. 0 indicates the lowest level of credibility, and 1 indicates the highest level of credibility.

In this paper, credibility is considered to be composed of 3 parts, Competency, Sincerity and Difficulty. Among them, Competency is the knowledge and skills possessed to accomplish this task. Integrity refers to the previous completion of the task. Difficulty of the task, on the other hand, refers to how easy or difficult it is to accomplish the task. Where competence is a more constant value that does not change much, the other 2 values change dynamically. Definitions regarding the relationship between trustworthiness and competence, integrity and difficulty are as follows:

Definition 2: Assuming that C is used to represent competence, S is used to represent integrity, and D is used to represent difficulty, the relationship between them is expressed as follows:

$$\begin{aligned} t &= \varphi(c,s,d) \\ S &\varnothing C \varnothing (1-D) \to T \end{aligned} \tag{2}$$

where, $c \in C$, $s \in S$, $d \in D$, $t \in T$. And $c,d,t \in (0,1)$. 0 denotes low competence, integrity and difficulty. 1 denotes high competence, integrity and difficulty.

Definition 3: For a given user or program x, its trustworthiness in handling task y can be expressed as:

$$t_x(y) = \varphi\left(c_x(y), s_x(y), d_x(y)\right) \tag{3}$$

And the following relationships exist:

$$\forall y \subset z, t_x(y) \le t_x(z) \tag{4}$$

Among them, operators have the following definitions:

Definition 4: Operator Definitions

Multiplication:

$$\varnothing = \varphi_{mul} = s \times c \times (1-d) \tag{5}$$

Min:

$$\varnothing = \varphi_{\min} = \min(s, c, (1-d)) \tag{6}$$

Combination:

$$\varnothing = \varphi_{com} = 1-(1-s)\times(1-c)\times d \tag{7}$$

They exist in the following relationship:

$$\sigma_{mul} \le \sigma_{\min} \le \sigma_{com} \tag{8}$$

Different operators can be used depending on the situation. In addition, a round value of plausibility is to be determined, as defined below:

Definition 5: Credibility Threshold (θ) is the minimum value of credibility. Less than this value is called untrustworthy and greater than or equal to this value is called trustworthy.

In this paper, the credibility of the corpus mainly includes two categories: the corpus credibility of the algorithm AT and the corpus credibility of the user UT.

The corpus trustworthiness of an algorithm refers to the degree of trustworthiness of a certain algorithm in textual information processing after its computation of the corpus. The trustworthiness of the results obtained by an algorithm after computation of a task also depends on three components: the capability of the algorithm, the degree of integrity and the difficulty of the task. In this case, the capability of the algorithm can be understood as the accuracy with which the algorithm handles the type of problem, the degree of integrity can be considered as 1, and the difficulty of the task needs to be determined based on what the algorithm handles.

The algorithm-based corpus credibility formula is as follows:

$$AT_x(y) = \sigma\left(c_x(y), 1, d_x(y)\right) = c_x(y) \times \left(1 - d_x(y)\right) \tag{9}$$

User-based corpus trustworthiness is the degree of trustworthiness of a corpus when it is proofread by the user. It is a reflection of the user's overall situation when completing similar tasks, and is related to that user's ability to perform the same type of task, how well it is completed, and the difficulty of the task. The credibility of user x in completing the y_ird category task for the jnd time is:

$$\begin{aligned} UT_x\left(y_{i,j}\right) &= \sigma\left(c_x\left(y_{ij-1}\right), s_x\left(y_{i,-1}\right), d_x\left(y_{i,-1}\right)\right) \\ &= 1-\left(1-c_x\left(y_{ij-1}\right)\right)\left(1-s_x\left(y_{i,-1}\right)\right)\right) d_x\left(y_{i,j-1}\right) \end{aligned} \tag{10}$$

where, $c_x\left(y_{ij-1}\right)$ refers to the capability obtained by the user x after completing the $j-1$th y_ith category task adjusted. The initial capability of the user is mainly classified into 4 categories, which are domain experts, specialized staff, registered users and general users, and the initial value $c_x\left(y_{i,j-1}\right)$ is taken as 0.9, 0.7, 0.5 and 0.3 according to the experiment, respectively. And $c_x\left(y_{ij}\right)$ is a dynamic value, which is automatically adjusted according to the credible situation of the user completing the category. The formula for the adjustment is shown below:

$$c_{x,j}\left(y_i\right)=\begin{cases}c_x\left(y_{i,-1}\right)+UT_x\left(y_{i,-1}\right)\times c_x\left(y_{i,-1}\right)\\ UT_x\left(y_{i,j-1}\right)\geq\theta\\ c_x\left(y_{i,j-1}\right)-\left(1-UT_x\left(y_{i,-1}\right)\right)\times c_x\left(y_{i,j-1}\right)/2\\ UT_x\left(y_{i,-1}\right)<\theta\end{cases} \tag{11}$$

The initial value $s_x\left(y_{i,0}\right)$ for each user is 0.5, a value that is automatically adjusted according to the user's credible completion of the class, mainly in the following ways:

$$s_x\left(y_{i,j}\right)=\begin{cases}s_x\left(y_{i,j-1}\right)+\dfrac{\sum s_z\left(y_{i,j-1}\right)}{n}\times s_x\left(y_{i,j-1}\right)\\ UT_x\left(y_{i,j-1}\right)\geq\theta\\ s_x\left(y_{i,j-1}\right)-\dfrac{\sum s_z\left(y_{i,j-1}\right)}{n}\times s_x\left(y_{i,j-1}\right)\\ UT_x\left(y_{i,j-1}\right)<\theta\end{cases} \tag{12}$$

where U denotes the set of all users participating in task y_i of $j-1$. n denotes the number of users in this set. And $d_x\left(y_{ij}\right)$ denotes the difficulty of this task, which is expressed by the following equation:

$$d_x\left(y_{i,j}\right)=1-\frac{1}{n\times l} \tag{13}$$

where, n denotes the number of users participating in task y_{ij}. l denotes the number of answers at the end of the task.

3.3. Syntactic structure classification based on BERT-Bi-LSTM-CRF

Based on the constructed large-scale Chinese corpus, this paper realizes the syntactic structure classification of Chinese corpus by combining the BERT model and Bi-LSTM-CRF model.

3.3.1. BERT model. The overall structure of the BERT model [13] is shown in Figure 2. Trm, i.e., denoting the Transformer model [14], suggests that the model is based on a bi-directional Transformer encoder, which is characterized by self-supervised learning of feature representations in a large-scale Chinese corpus, and the use of BERT features as high-quality word embeddings for natural language processing tasks.

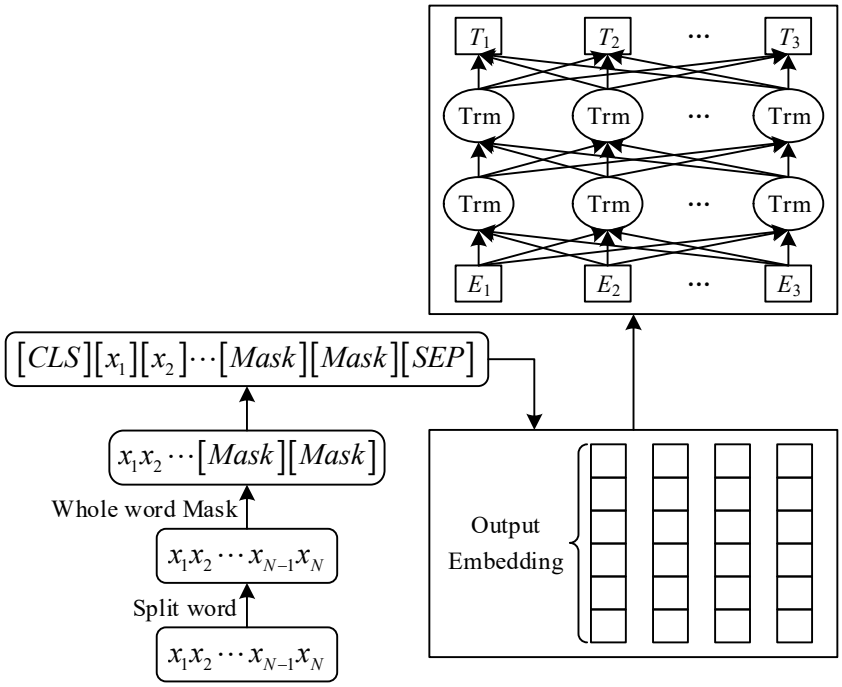

Fig. 2. BERT masked language model

BERT pretraining is divided into masked word prediction task (MLM) and next sentence prediction task (NSP). In this paper, we use masked word pre-training, when generating training samples, 15% of the subwords in the sentence will be randomly masked, the masking methods include replacing with fixed identifiers, keeping the original word and randomly replacing with other words, and generating word vectors which are then inputted into BERT to extract features. The word vector is generated and fed into BERT to extract features. Only the masked part is predicted based on the contextual relationship during the specific training. In the BERT dictionary, special identifiers are added. [CLS] is a sentence-initial identifier. [SEP] is a separator, which is used to separate 2 independent sentences. [UNK] is an unknown identifier. [MASK] is a masking identifier, with an 80% probability of being masked by [MASK] in 15% of the sequence for a random masking strategy, another 10% probability of replacing it with a word in the text sequence, and a 10% probability of leaving it unchanged.

3.3.2. Bi-LSTM-CRF modeling. Bi-LSTM adopts a bidirectional long and short-term memory neural network, which is good at discovering the correlation relationship between characters, capturing the information of the long context sequence of Chinese corpus, and possessing the ability of neural network to fit the nonlinearities, and utilizes gated units to realize long-term memory, which solves the problem of gradient disappearance or gradient explosion during the training of recurrent neural network [15]. LSTM controls the unit state through input gates, output gates, and oblivion gates. The input gate accepts the saved information at the current moment, the output gate controls the process from the current state to the LSTM output, and the forgetting gate determines the information that can be retained from moment $t-1$ to moment t in the cell state, see Eqs. (14) to (18):

$$i_t = \sigma\left(W_i \cdot \left[h_{t-1}, x_t\right] + b_i\right) \tag{14}$$

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \tag{15}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tanh\left(W_c \cdot \left[h_{t-1}, x_t\right] + b_c\right) \tag{16}$$

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right) \tag{17}$$

$$h_t = o_t \cdot \tanh\left(c_t\right) \tag{18}$$

where: W_i, W_f, W_c are the weight matrices of the input gate, forget gate, and output gate, respectively. b_i, b_f, b_c are the deviation terms. The current input value and the value of the forgetting gate are obtained by the output h_{t-1} and the current input x_t at the moment of $t-1$, and then the unit state c_t at the moment of t is obtained according to the unit state c_{t-1} and the current input value at the moment of $t-1$, so as to realize the combination of the current memory and the long-term memory, i.e., the long-time sequence relationship. c_t The t moment output h_t obtained by multiplying with the value of the output gate after transforming by the $\tanh$ function.

However, Bi-LSTM only outputs the predicted labels based on the maximum probability, and the outputs are not affected by each other, which leads to sequences such as "B-PER" followed by "I-ORG". CRF model is a conditional probability distribution model underlying natural language processing, defined as a probability problem to find the distribution of the output variable given the input random variables [16]. Let X and Y be random variables and $P(Y \mid X)$ be the conditional distribution of Y under X conditions. Where if Y is a Markov random field consisting of undirected graphs, then the corresponding $P(Y \mid X)$ is called a conditional random field. The basic feature of "conditional" is to find the probability of state sequence Y based on the observation sequence X. The advantage of CRF is to learn the implicit conditions between states, to consider the local features of sentences more, to obtain the optimal sequence by proximity labeling, and to make up for the shortcomings of Bi-LSTM. Therefore, the combination of Bi-LSTM and CRF model is considered, which can maintain long-term memory and consider local dependencies, and the structure of Bi-LSTM-CRF model is shown in Fig. 3.

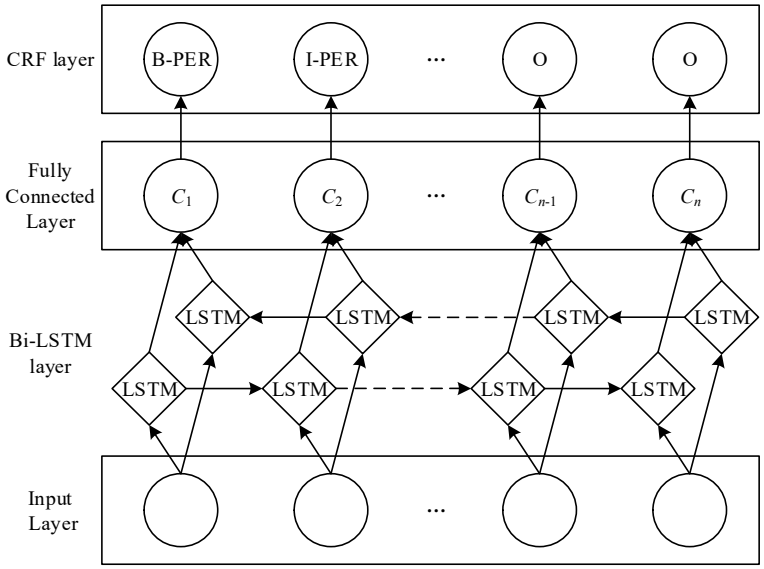

Fig. 3. The model structure of Bi-LSTM-CRF

If the labeling sequence of a particular sentence x is $y = \left(y_1, y_2, \cdots, y_n\right)$, the score of the labeling sequence y of sentence x is obtained under the Bi-LSTM-CRF model:

$$S(x, y) = \sum_{i=1}^{n} P_{i,yi} + \sum_{i=1}^{n+1} Q_{yi-1,yi} \tag{19}$$

where: P_i, y_i is the output score matrix of Bi-LSTM. Q_{yi-1}, yi is the transfer score from the $i-1$th label to the ith label. The scores are determined by the output of the Bi-LSTM layer and the transfer

matrix of the CRF, respectively. The labeling result probability, see equation (20), where y' is the true sequence. The probabilities are taken logarithmically to obtain the likelihood function solution, see equation (21). Finally, the objective of the likelihood function is to output the most satisfactory score sequence as the predicted sequence, see equation (22). Namely:

$$P(y \mid x) = \frac{\exp(S(x, y))}{\sum_{y \in YX} \exp\left(S\left(x, y'\right)\right)} \tag{20}$$

$$\lg(P(y \mid x)) = S(x, y) - \lg\left(\sum_{y' \in XX} \exp\left(S\left(x, y'\right)\right)\right) \tag{21}$$

$$y^* = \arg\max_{y \in \gamma X} S\left(x, y'\right) \tag{22}$$

4. Experimental results and analysis

4.1. Evaluation indicators

The experiments in this paper are supervised learning carried out based on labeled data, and both the training and test sets are sequence data, so the sequence label-based evaluation method is used to evaluate the model classification effect from three indicators, namely, Precision (P), Recall (R), and the reconciled mean F-measure (F). The calculation methods of the three are shown in Eqs. (23) to (25):

$$\Pr ecision\left(P\right) = \frac{A}{A + B} \times 100\% \tag{23}$$

$$\mathrm{Re}\, call\left(R\right) = \frac{A}{A + C} \times 100\% \tag{24}$$

$$F - measure\left(F\right) = \frac{2 \times P \times R}{P + R} \times 100\% \tag{25}$$

Where A denotes the number of syntactic structures accurately classified by the model, B denotes the number of syntactic structures present in the manually labeled sequences, and C denotes the number of syntactic structures present in the sequences labeled by the model. From the principle of the final formula, the reconciled mean F-measure (F) integrates the accuracy and recall, which is more objective and reasonable for the evaluation of the final classification effect of the model. At the same time, when counting the values of A, B and C, it is important to adhere to the overall principle, i.e., a syntactic structure is considered to be classified accurately only if it is correctly classified by the overall F, and to avoid counting the labeled accuracy rate, recall rate and the reconciliation mean individually.

4.2. Large Scale Chinese Syntactic Structure Classification Corpus Construction

In order to fully prove the performance of the classification model, this paper classifies syntactic structures from two aspects, modern Chinese and ancient Chinese, respectively. Therefore, the corpus of this paper has two parts: first, the structured Chinese Treebank of Tsinghua University. The second is the corpus of Chinese literary texts using the corpus construction model. In view of the differences in syntactic structure between modern Chinese and ancient Chinese, this paper only selects two kinds of verbal phrases, namely, stating object and stating complement, for syntactic structure labeling and classification.

4.2.1. Tsinghua Chinese Treebank. The Tsinghua Chinese Treebank is a large-scale, high-quality Chinese syntactic treebank corpus containing one million Chinese characters, extracted from a large-scale Chinese corpus that has undergone lexical segmentation and lexical annotation, and has been

formed after automatic break direction, automatic syntactic analysis and manual proofreading. The basic statistics of the Tsinghua Chinese Treebank selected for this paper are shown in Table 1.

Table 1. Basic statistics of Tsinghua Chinese Library

Literary form	Number of files	Sentence count	Lexical number	Number of Chinese characters	Average word (word/sentence)
News	154	6877	173942	246757	25.29
Literature	139	16335	340208	415040	20.83
Apply	195	3169	66586	97924	21.01
Science	15	5589	158780	240289	28.41
Total	503	31970	739516	1000010	23.13

The tokens in the treebank mainly contain lexical tokens and syntactic relation tokens. Since the ultimate goal of this experiment is to identify the categories of verb phrase structure, this paper only retains 21 major categories of lexicality in the original corpus preprocessing of the Tsinghua Chinese Treebank as shown in Table 2, and at the same time, punctuation is uniformly labeled as "w".

Table 2. Tsinghua Chinese tree database original corpus parts of speech

Part-of-speech marking	The character of a certain word	Part-of-speech marking	The character of a certain word
a	Adjective	o	Onomatopoeic word
b	Distinguishing word	p	Preposition
c	Conjunction	q	Quantifier
d	Adverb	r	Pronoun
e	Interjection	s	Onomatopoeic word
f	Noun of locality	t	Time words
h	Prefix	u	Auxiliary word
i	Affix	v	Verb
k	Subsequent element	y	Modal particle
m	Numeral	z	Adverbial
n	Noun	w	Punctuation

There are 27 syntactic relation markers in the treebank. In this paper, we use the structurally adjusted Tsinghua Chinese Treebank, i.e., the Chinese corpus that retains the structure of verbal phrases is divided into two groups, which retains the stating-object (movable-object) structure (PO) and stating-supplement (movable-supplement) structure (SB), and takes the minimal structure in the nested structure. The distribution of the minimal structures of the two types of verbal phrases in the corpus is shown in Table 3.

Table 3. Distribution of minimum structure of verbal phrase in Tsinghua Chinese tree library

Structure category	Structural quantity	Average word number	Total number of words	Proportion
Complement structure (SB)	9142	2.37	698574	3.10%
Predicate-object construction (PO)	61857	2.75	698574	24.35%

4.2.2. Corpus of Chinese Texts:

1) Introduction of Canonical Books. The six literary canons used in this paper for automatic categorization of syntactic structures of Chinese verbal phrases are Shangshu, Zuozhuan, Tao Te Ching, Lunyin, Zhan Guo Ce, and Li Ji.

2) Corpus Source and Acquisition. Based on the research of existing Chinese electronic corpus resources, this paper finally selects the data source of "Chinese Philosophical Book Electronic Program", which is an online open electronic library and literature retrieval system. In this paper, the data of ancient Chinese canonical books mainly come from the corpus of ancient Chinese canonical books in this database. In this study, the corpus of ancient books in modern language comes from the Ancient Poetry Website, which uses a crawler to capture the vernacular texts of six canonical books, and together with the above-mentioned corpus of ancient books in ancient Chinese, it constitutes a corpus of Chinese canonical books in two languages.

3) Original corpus processing and corpus construction. For the ancient modern text corpus, since it is the same as the Tsinghua Chinese Treebank for Chinese modern text, it is possible to realize the recognition of participle and lexicality with the help of the well-labeled Tsinghua Chinese Treebank. In the task of word segmentation, we use the sequence labeling of {B, I, E, S}, which represents the beginning, middle, end and single words respectively. After the corpus is processed into the target format, the word segmentation model trained by Tsinghua Chinese Treebank is used for word segmentation. After the completion of the lexical annotation, according to the annotation sequence, the corpus is processed into the form of word as a unit, which can be used for subsequent lexical annotation and verbal phrase structure annotation.

For the original text corpus of ancient books, due to the large gap between the modern text corpus of Tsinghua Chinese Treebank and it, this paper adopts a combination of machine annotation and manual proofreading in the process of the annotation of participle and lexicality.

For the sake of efficiency, the CRF model is used for the training of the integrated labeling model of ancient Chinese participles and lexemes. Randomly select 1/10 of the whole corpus and divide this 1/10 into training set and test set according to the ratio of 9:1, in which the size of the training set is 1652KB.The best results of the training of the integrated model of ancient text participle lexicality are shown in Table 4. The overall accuracy, recall and F1 value of ancient text participle lexicalization reached 84.34%, 84.86% and 84.60%, respectively, and the model training is more effective.

Table 4. Training results of the integrated model of lexical segmentation in ancient Chinese

Tag	Implication	Precision	Recall	F1 value	Number
Total		84.34%	84.86%	84.60%	20902
a	Adjective	52.57%	32.33%	40.04%	271
c	Conjunction	87.85%	87.06%	87.45%	743
d	Adverb	84.55%	84.71%	84.63%	1159
f	Noun of locality	64.29%	55.06%	59.32%	94
gv	Archaic verb	0.00%	0.00%	0	0
j	Conversion words	61.03%	47.52%	53.43%	16
m	Numeral	85.75%	87.83%	86.78%	321
n	Noun	69.43%	73.69%	71.50%	4297
nr	Name of a Person	75.67%	68.31%	71.80%	915
ns	Place Name	72.90%	57.93%	64.56%	226
p	Preposition	85.54%	85.37%	85.45%	578
q	Quantifier	83.71%	65.19%	73.30%	20
r	Pronoun	89.51%	90.29%	89.90%	1342
s	Onomatopoeic word	0.00%	0.00%	0	0
t	Time noun	85.66%	71.16%	77.74%	152
u	Auxiliary word	86.28%	76.97%	81.36%	446
v	Verb	84.18%	86.40%	85.28%	5085
w	Punctuation	99.87%	100.00%	99.93%	4679
y	Modal particle	87.29%	92.49%	89.81%	558

Then, the trained model is used to automatically label the ancient text corpus of the six canonical books with participles and lexemes. In order to further ensure the accuracy of the annotation results, manual proofreading is also required on the basis of machine annotation. The manual proofreading platform chosen in this paper is Brat, a text annotation tool under Linux system.

4.2.3. Calculation of corpus credibility. Separate experiments were conducted on a corpus of Chinese canonical syntactic structure categorization using the plausibility calculation method. Using plausibility it is possible to pick out 10.42% of all corpora that need proofreading, and among these untrustworthy corpora, 75.37% of the misclassified emails are labeled as untrustworthy. Through manual proofreading as well as credibility calculation, the average classification correctness of the final Chinese canonical corpus is increased from 94.21% to 99.06%, which is an improvement of 4.85%. The experiment proves the effectiveness of the credibility calculation.

4.3. Chinese Segmentation and Lexical Pre-Labeling Model Training

Since Chinese does not naturally come with participle features, thus before realizing the automatic classification of verbal syntactic structure of Chinese canonical books, it is necessary to first realize its automatic participle, and then realize lexical annotation based on the results of automatic participle. In the training process of the lexical model, this paper adopts the BERT-Bi-LSTM-CRF model as Baseline, and extracts all the words according to "/" in the Tsinghua Chinese Treebank, and then labels each word as {B, I, E, S} according to its position in the vocabulary. The results of the ten-fold cross-validation of the automatic word segmentation model are shown in Table 5. The mean value of F-measure for ten experiments is 92.09%, which is relatively reliable, and an optimal participle model with 92.41% accuracy and 93.18% recall can be obtained from Experiment 6.

Table 5. Chinese automatic word segmentation ten-fold cross verification results

Experiment number	The number of the training corpus	The number of the test corpus	P	R	F
1	1-9	0	91.41%	92.83%	92.11%
2	0,2-9	1	91.92%	92.49%	92.20%
3	0-1,3-9	2	91.41%	91.53%	91.47%
4	0-2,4-9	3	91.51%	92.46%	91.98%
5	0-3,5-9	4	92.02%	92.96%	92.49%
6	0-4,6-9	5	92.41%	93.18%	92.79%
7	0-5,7-9	6	91.22%	91.14%	91.18%
8	0-6,8-9	7	93.64%	91.56%	92.59%
9	0-7,9	8	91.07%	92.35%	91.71%
10	0-8	9	92.53%	92.29%	92.41%
Mean value			91.91%	92.28%	92.09%

Then, this paper then starts from Tsinghua Chinese Treebank, extracts each word and the corresponding lexical property with it, saves it as the format for machine training, and carries out the training of Chinese lexical automatic classification model, and the results of the ten-fold cross-validation of the lexical automatic classification model are shown in Table 6. According to the information in the table, the average value of F-measure of the lexical model is 89.13%, and the optimal experiment is experiment 8, with an F-measure of 90.75%, which is also at a high level.

Table 6. Ten-fold cross validation of automatic classification of Chinese parts of speech

Experiment number	The number of the training corpus	The number of the test corpus	P	R	F
1	1-9	0	90.34%	90.25%	90.29%
2	0,2-9	1	91.18%	89.76%	90.46%
3	0-1,3-9	2	90.63%	89.05%	89.83%
4	0-2,4-9	3	88.27%	88.17%	88.22%
5	0-3,5-9	4	88.35%	89.08%	88.71%
6	0-4,6-9	5	89.89%	90.77%	90.33%
7	0-5,7-9	6	89.69%	91.54%	90.61%
8	0-6,8-9	7	91.83%	89.69%	90.75%
9	0-7,9	8	88.22%	90.03%	89.12%
10	0-8	9	83.05%	82.83%	82.94%
Mean value			89.15%	89.12%	89.13%

4.4. Comparative Experiments on Classification of Chinese Syntactic Structures

In order to verify the effectiveness of the proposed large-scale corpus-based Chinese syntactic structure classification model BERT-Bi-LSTM-CRF, this paper compares it with the BERT model and the Bi-LSTM-CRF model for Chinese syntactic structure classification.

4.4.1. BERT model classification results. Chinese syntactic structure classification experiments are divided into three groups, the first group of experiments is based on the BERT pre-training model classification experiments, the data object of the experimental group is comprehensive data, i.e., the data include the “statement of the complement structure”, “statement of the object structure” and other data structures, and all the labeling results are uniformly calculated, and the BERT model ten-fold experimental classification results are shown in Table 7. The results of all labels are calculated uniformly, and the classification results of the ten-fold experiment of the BERT model are shown in Table 7.

As can be seen from Table 7, the results of the ten-fold cross-tabulation experiments are relatively stable, with no large or small values, and the ranges of the accuracy, recall, and reconciliation mean are [35.50%,39.85%], [33.57%,38.04%], and [35.15%,37.91%], respectively, and the corresponding arithmetic means are 37.96%, 35.64%, and 36.73%, respectively. 36.73%. From the recognition results, the overall recognition effect is not good enough, the optimal F-measure value is only 37.91%, and the overall recall rate is slightly lower than the accuracy rate (35.64%<37.96%). This is because the deep learning model essentially calculates the conditional probability of words, and if the frequency of words is too low, forming too few sentences and contexts, the features learned by the model will also be too low, which leads to a lower recognition effect of various types of Chinese syntactic structures in the test set.

Table 7. BERT model ten-fold experiment classification results

Experimental serial number	Precision	Recall	F-measure
1	39.36%	33.72%	36.32%
2	38.56%	34.23%	36.27%
3	35.50%	35.34%	35.42%
4	36.74%	38.04%	37.38%
5	38.51%	37.30%	37.90%
6	38.41%	37.43%	37.91%
7	37.61%	35.21%	36.37%
8	36.89%	33.57%	35.15%
9	38.15%	37.54%	37.84%
10	39.85%	34.01%	36.70%
Arithmetic mean	37.96%	35.64%	36.73%

4.4.2. Bi-LSTM-CRF model classification results. The second group of experiments is the classification experiments based on the Bi-LSTM-CRF pre-training model, the experimental data objects of this group are the same as the first group of experiments, which also include the "complementary structure", "binning structure" and other data structures, and the experimental results of the Bi-LSTM-CRF model are shown in Table 8. The experimental results of the CRF model are shown in Table 8.

The recognition results of the ten-fold crossover experiments based on the Bi-LSTM-CRF pre-trained model are also relatively stable, and there are no cases of extreme data. The accuracy, recall, and reconciliation averages are within the intervals of [44.35%,48.40%], [43.15%,48.83%], and [43.89%,47.97%], respectively, and the corresponding algorithm averages are 45.99%, 46.20%, and 46.06%, respectively.

From the final results, the overall recognition effect is not high, with the optimal tuning and the average value of F-measure being only 47.97%, but the optimal values of Precision, Recall, and F-measure are improved by 8.55%, 10.79%, and 10.06%, respectively, and the average value of F-measure is improved by 8.03%, 10.56% and 9.33%. From the classification results, similar to the classification results of the BERT pre-trained model, the syntactic structure classification of short texts is more excellent, and the classification of long texts and texts with complex structures is less effective.

Table 8. Bi-LSTM-CRF model ten-fold experiment classification results

Experimental serial number	Precision	Recall	F-measure
1	47.32%	46.59%	46.95%
2	45.15%	43.31%	44.21%
3	44.35%	48.73%	46.44%
4	44.99%	48.25%	46.56%
5	47.14%	48.83%	47.97%
6	48.40%	44.01%	46.10%
7	45.48%	44.54%	45.01%
8	44.66%	43.15%	43.89%
9	46.71%	46.87%	46.79%
10	45.66%	47.71%	46.66%
Arithmetic mean	45.99%	46.20%	46.06%

4.4.3. BERT-Bi-LSTM-CRF model classification results. The third group of experiments is the classification experiments based on the pre-trained model of BERT-Bi-LSTM-CRF, which is divided into three categories, namely BERT-Bi-LSTM-CRF1, BERT-Bi-LSTM-CRF2, and BERT-Bi-LSTM-CRF3. The first two types of experimental data objects are the "complementary structure" and the "object structure" respectively, while the third type of experimental data objects is the data set of all structures. Ten-fold crossover experiments were conducted for each of the three groups of models, totaling 30 rounds of experiments. The experimental results are shown in Tables 9, 10 and 11, respectively.

The data in Table 9 show that the accuracy P, recall R, and reconciliation mean F of the BERT-Bi-LSTM-CRF1 model lie between [50.33%,56.41%], [50.53%,54.47%], and [51.52%,54.57%], respectively, and their arithmetic means are 53.66%, 52.21%, and 52.90%, respectively.

Table 9. BERT-Bi-LSTM-CRF1 model ten-fold experiment classification results

Experimental serial number	Precision (P)	Recall (R)	F-measure (F)
1	50.33%	53.65%	51.94%
2	55.28%	52.89%	54.06%
3	55.08%	51.94%	53.46%
4	53.26%	53.23%	53.24%
5	53.49%	54.47%	53.98%
6	53.98%	50.54%	52.20%
7	53.40%	51.27%	52.31%
8	53.06%	50.53%	51.76%
9	56.41%	52.84%	54.57%
10	52.29%	50.78%	51.52%
Arithmetic mean	53.66%	52.21%	52.90%

As can be seen from Table 10, the optimal values of accuracy P, recall R, and reconciliation mean F of the BERT-Bi-LSTM-CRF2 model are 54.99%, 52.58%, and 53.48%, respectively, and the arithmetic means are 53.24%, 51.11%, and 52.14%, respectively.

Table 10. BERT-Bi-LSTM-CRF2 model ten-fold experiment classification results

Experimental serial number	Precision (P)	Recall (R)	F-measure (F)
1	54.63%	52.37%	53.48%
2	54.02%	51.75%	52.86%
3	53.62%	50.58%	52.06%
4	53.90%	49.72%	51.73%
5	52.56%	51.44%	51.99%
6	51.80%	49.52%	50.63%
7	53.51%	51.89%	52.69%
8	52.98%	52.58%	52.78%
9	54.99%	49.76%	52.24%
10	50.34%	51.51%	50.92%
Arithmetic mean	53.24%	51.11%	52.14%

The data in Table 11 show that the Precision, Recall, and F-measure of the BERT-Bi-LSTM-CRF3 model are within the intervals [48.80%,52.63%], [47.07%,50.76%], and [48.44%,51.65%], respectively, and the corresponding arithmetic means are 51.02%, 48.92% and 49.94%, respectively.

Table 11. BERT-Bi-LSTM-CRF3 model ten-fold experiment classification results

Experimental serial number	Precision (P)	Recall (R)	F-measure (F)
1	51.58%	50.76%	51.17%
2	48.80%	48.29%	48.54%
3	49.89%	47.07%	48.44%
4	49.83%	49.01%	49.42%
5	52.63%	50.71%	51.65%
6	51.55%	49.86%	50.69%
7	50.90%	50.09%	50.49%
8	50.42%	48.04%	49.20%
9	52.15%	48.28%	50.14%
10	52.41%	47.09%	49.61%
Arithmetic mean	51.02%	48.92%	49.94%

The optimal and mean enhancements of the BERT-Bi-LSTM-CRF model compared to the BERT model, Bi-LSTM-CRF model, BERT-Bi-LSTM-CRF1 model and BERT-Bi-LSTM-CRF2 model are shown in Table 12.

The classification effect of the BERT-Bi-LSTM-CRF3 model trained on the data of "complement structure + object structure + other structure" in the data of "Tsinghua Chinese Tree Library" is better than that of the BERT model, which is not much different from the Bi-LSTM-CRF model but slightly higher than that of Bi-LSTM-CRF, but lower than that of the BERT-Bi-LSTM-CRF1 model and the BERT-Bi-LSTM-CRF2 model. Analyzing the experimental output results, the following conclusion can be drawn: the continued pre-training model that fuses multiple complex features performs better than the most original model. At the same time, its performance is lower than that of the continued pre-training model that fuses single feature information, especially when classifying a single syntactic structure, too many irrelevant features are added, which may produce negative optimization results.

Table 12. BERT-Bi-LSTM-CRF3 model lifting amplitude

	Optimal value			Arithmetic mean		
Comparison model	Precision	Recall	F-measure	Precision	Recall	F-measure
BERT	12.78%	12.72%	13.74%	13.06%	13.28%	13.21%
Bi-LSTM-CRF	4.23%	1.93%	3.68%	5.03%	2.72%	3.88%
BERT-Bi-LSTM-CRF1	-3.78%	-3.71%	-2.92%	-2.64%	-3.29%	-2.96%
BERT-Bi-LSTM-CRF2	-2.36%	-1.82%	-1.83%	-2.22%	-2.19%	-2.20%

4.5. Data Labeling Validation

Continuing the pre-training experiments, we obtain the autonomous pre-trained models BERT-Bi-LSTM-CRF1 and BERT-Bi-LSTM-CRF2 for the automatic recognition of the "Complementary Structures" and "Bin Structures", and select the models with the best F-values in the experiments to carry out labeling experiments on the unlabeled data set. The models that achieved the optimal F-values in the experiments were selected to perform labeling experiments on the unlabeled data sets. The experiments are comparison experiments, respectively, 300 unlabeled data are randomly obtained from the Chinese canonical corpus data, the first group of data are all labeled manually, and the second group of data are first labeled with the model, and then reviewed using manual review. After all the labeling is completed, the labeling results are tested and counted, and the results of the data labeling comparison experiment are obtained as shown in Table 13.

As can be seen from Table 13, when the syntactic structure annotation is performed on the original data, the combination of model annotation and manual annotation has almost the same accuracy as that of manual annotation (only a difference of 0.66%), but the average time of annotation has been reduced by 37.04%, which reduces the workload of the annotators and improves the efficiency of annotation. And with the further expansion of the number of annotations, the attention of the annotators decreases, the efficiency will be further reduced, and the "model + manual" annotation method will highlight the advantages. This verifies the effectiveness and practicality of the BERT-Bi-LSTM-CRF model proposed in this paper in automatic classification of Chinese syntactic structures.

Table 13. Data annotation comparison experiment

	Manual labeling	"Model + manual" labeling
Marking quantity	300	300
Average time spent	162 minutes	102 minutes
Correct number	292	290
Number of errors	8	10
Accuracy	97.33%	96.67%

5. Conclusion

This paper proposes a grid-based large-scale corpus construction and distribution model, and a Chinese syntactic structure classification model BERT-Bi-LSTM-CRF, which realizes the classification of Chinese syntactic structures using a large-scale corpus.

Through manual proofreading as well as calculating the credibility of the corpus, the average correct rate of syntactic structure classification of the Chinese canonical corpus constructed in this paper is increased from 94.21% to 99.06%, which is an improvement of 4.85%, proving the effectiveness of the proposed credibility calculation method. And through the training of Chinese participle and lexical pre-annotation model, this paper obtains the best participle model for the constructed Chinese large-scale corpus.

The BERT-Bi-LSTM-CRF1 model with the experimental data of "declarative structure" has the optimal values of 56.41%, 54.47%, and 54.57% for accuracy, recall, and reconciliation mean, respectively, and the corresponding arithmetic means are 53.66%, 52.21%, and 52.93% for the BERT

model, Bi-LSTM-CRF model, and the BERT-Bi-LSTM-CRF model with "declarative structure" and all syntactic structures, respectively. 52.93%, which are better than the BERT model, the Bi-LSTM-CRF model, and the BERT-Bi-LSTM-CRF2 and BERT-Bi-LSTM-CRF3 models with the experimental data of "object structure" and all syntactic structures, respectively. This shows that the performance of the syntactic structure classification model fusing multiple complex features is better than the original model, while its performance is lower than that of the syntactic structure classification model fusing single feature information. Adding too many irrelevant features may produce negative optimization results.

In addition, when the syntactic structure annotation is performed on the original data, the combination of BERT-Bi-LSTM-CRF model annotation and manual annotation reduces the average time spent on annotation by 37.04% and improves the efficiency of annotation while the accuracy is almost the same as that of manual annotation. This indicates that the BERT-Bi-LSTM-CRF model proposed in this paper is suitable for the task of automatic classification of Chinese syntactic structures.

References

[1] Sun, Y. (2018). A Cartographic Analysis of the Syntactic Structure of Mandarin. Studies in Chinese Linguistics, 39(2), 127-154.

[2] WANG, Y., ZHANG, L., CUI, N., & WU, Y. (2023). The role of syntactic structure and verb overlap in spoken sentence production of 4-to 6-year-olds: evidence from syntactic priming in mandarin. Acta Psychologica Sinica, 55(10), 1608.

[3] Dai, W. (2021). On the Syntactic Structure of Chinese Ambiguity Sentences. Open Access Library Journal, 8(10), 1-10.

[4] Niu, R., & Osborne, T. (2019). Chunks are components: A dependency grammar approach to the syntactic structure of Mandarin. Lingua, 224, 60-83.

[5] Yang, S., Cai, Y., Xie, W., & Jiang, M. (2021). Semantic and syntactic processing during comprehension: ERP evidence from Chinese QING structure. Frontiers in Human Neuroscience, 15, 701923.

[6] Chen, L., Gao, C., Li, Z., Zaccarella, E., Friederici, A. D., & Feng, L. (2023). Frontotemporal effective connectivity revealed a language-general syntactic network for Mandarin Chinese. Journal of Neurolinguistics, 66, 101127.

[7] Li, Y., Szmrecsanyi, B., & Zhang, W. (2023). The theme-recipient alternation in Chinese: tracking syntactic variation across seven centuries. Corpus Linguistics and Linguistic Theory, 19(2), 207-235.

[8] Yu, Q. (2021). An organic syntactic complexity measure for the Chinese language: The TC-unit. Applied Linguistics, 42(1), 60-92.

[9] Liao, W. W. R., & Lin, T. H. J. (2019). Syntactic structures of Mandarin purposives. Linguistics, 57(1), 87-126.

[10] Skvortsov, A. V., & Malykh, O. A. (2020). Teaching Russian students to analyze the syntactic structure of sentences in modern Chinese (Mandarin). Vestnik of Samara State Technical University Psychological and Pedagogical Sciences, 17(2), 153-172.

[11] Lili, G. (2017). A Brief Contrastive Analysis of English and Chinese Syntactic Structure. In Asia International Symposium on Language, Literature and Translation (p. 219).

[12] Xiaona, W. E. I., & Min, Y. A. N. (2022). A Study on the Effects of Syntactic Structure on the Readability of Chinese Texts. Journal of Teacher Education, 9(6), 81-87.

[13] Mandana Sadat Ghafourian, Sara Tarkiani, Mobina Ghajar, Mohamad Chavooshi, Hossein Khormaei & Amin Ramezani. (2025). Optimizing warfarin dosing in diabetic patients through BERT model and machine learning techniques. Computers in Biology and Medicine109755-109755.

[14] Jing Rong, Cong Wang, Qiuzhan Zhou, Yunxue He & Huinan Wu. (2025). Enhancing non-intrusive load monitoring through transfer learning with transformer models. Energy & Buildings115334-115334.

[15] Zhongying Wang, James L. Crooks, Elizabeth Anne Regan & Morteza Karimzadeh. (2025). High-Resolution Estimation of Daily PM2.5 Levels in the Contiguous US Using Bi-LSTM with Attention. Remote Sensing(1),126-126.

[16] Tianshu Fang, Yuanyuan Yang & Lixin Zhou. (2024). Enhanced Precision in Chinese Medical Text Mining Using the ALBERT+Bi-LSTM+CRF Model. Applied Sciences(17),7999-7999.