Research on Named Entity Recognition Methods in Chinese Forest Disease Texts

Qi Wang and Xiyou Su *

School of Information Science and Technology, Beijing Forestry University, Beijing 100083, China; wangqi77@bjfu.edu.cn
* Correspondence: suxy@bjfu.edu.cn

Abstract: Named entity recognition of forest diseases plays a key role in knowledge extraction in the field of forestry. The aim of this paper is to propose a named entity recognition method based on multi-feature embedding, a transformer encoder, a bi-gated recurrent unit (BiGRU), and conditional random fields (CRF). According to the characteristics of the forest disease corpus, several features are introduced here to improve the method’s accuracy. In this paper, we analyze the characteristics of forest disease texts; carry out pre-processing, labeling, and extraction of multiple features; and construct forest disease texts. In the input representation layer, the method integrates multi-features, such as characters, radicals, word boundaries, and parts of speech. Then, implicit features (e.g., sentence context features) are captured through the transformer’s encoding layer. The obtained features are transmitted to the BiGRU layer for further deep feature extraction. Finally, the CRF model is used to learn constraints and output the optimal annotation of disease names, damage sites, and drug entities in the forest disease texts. The experimental results on the self-built data set of forest disease texts show that the precision of the proposed method for entity recognition reached more than 93%, indicating that it can effectively solve the task of named entity recognition in forest disease texts.

Keywords: disease; named entity recognition; multi-feature; transformer; bi-gated recurrent unit; CRF

1. Introduction

Named entity recognition is a core task in the field of natural language processing. Its goal is to extract specific types of entities from text, such as people’s names, place names, and organization names [1]. It plays a key role in the research of knowledge map construction, automatic question answering, and network searches. There are three basic kinds of named entity recognition methods: rule-based methods, statistical machine learning methods, and deep learning methods. With the development of deep learning technologies, such as long short-term memory (LSTM) and transformer, the performance of named entity recognition methods for the general domain has been greatly improved [2]. In particular, the transformer model processes the characters in the sequence in parallel, and uses the self-attention mechanism to directly model the relationship between the characters in the sentence instead of using the original cycle and convolution structure. It has better computational performance than RNNs and CNNs in tasks such as machine translation and named entity recognition. However, as the data in different fields have unique characteristics, and as there may be a lack of large-scale annotation data, research on named entity recognition in the general domain cannot be well-migrated to the proprietary domain. Therefore, scholars have carried out much exploratory research on named entity recognition in proprietary fields such as electronic medical records [3], bridge engineering [4], and military software detection [5].

Forest reserves are one of the most important resources for economic and social development, bringing rich economic, ecological, and social benefits. Forest diseases may
destroy large forests and cause poor tree growth, decline in yield and quality, and even the death of trees, resulting in economic losses and threatening the healthy development of the ecological environment. In history, serious diseases have been treated continuously. Their impact increases with the increase in plantation area. For example, Valsa sordida and poplar canker in northern China and early defoliation disease of Larch and Lecanosticta acicola in southern China pose serious threats to forestry production and ecology. Due to the concealment of forest diseases, it is difficult to distinguish and control them. Research on forest diseases and how to control these have always been a focus of attention. With the development of forestry information, information technology has become more widely used in the field of forest diseases, and substantial text data related to forest diseases have been accumulated, mostly stored in an unstructured form. Research on named entity recognition technology for forest disease-related texts can lay the foundation for the scientific construction of high-quality forest disease intelligent question-and-answer systems, recommendation systems, intelligent searches, and other downstream applications. The text in the field of forest diseases lacks large-scale annotation data sets. Entities have nested relationships. In addition, there are numerous domain-specific entity concepts and rare words in the text, such as diseases, damage parts, and so on. Named entity recognition of forest disease texts still needs to be explored in depth. For forestry texts, researchers have successively proposed a rule-based method [6], BCC-P method [7], etc., but these methods do not fully consider the characteristics of domain texts. Moreover, there is less research on named entity recognition of Chinese forest disease texts. Previous research has achieved better performance of named entity recognition in proprietary fields by multi-feature embedding. For example, in the field of Chinese electronic medical records, researchers embedded character feature and glyph feature as input, and had good results in the self-built data set [3].

In this study, we examine the named entity recognition of forest disease texts. The research mainly includes two parts: multi-feature embedding and the named entity recognition method for forest disease texts based on Transformer-BiGRU-CRF. Firstly, a corpus of named entities of forest diseases is constructed through pre-processing and tagging. Then, we analyze the characteristics of the texts. For example, according to the text features, many diseases and fungicides have specific radicals; a large number of entities do not have obvious boundary characteristics; and there are certain rules in the part-of-speech distribution of some entities, word boundary features, partial radical features, and part-of-speech features; three kinds of artificial features and word vector splicing are used as the input. Secondly, we combine the transformer encoder with relative position information to model the relationship between characters, taking advantage of the bidirectional feature extraction ability of BiGRU. A named entity recognition method for forest disease text based on Transformer-BiGRU-CRF is proposed. A corpus of named entities of forest diseases is constructed through pre-processing, tagging, and multi-feature extraction. Under the two scenarios of adding multiple features and not adding multiple features as inputs, a comparative experiment with the current mainstream models in named entity recognition is carried out. Finally, the optimal annotation of disease names, damage sites, and agent entities in forest disease text is achieved. On the basis of considering the multi-features of the data set texts as the model input, this method makes full use of transformer abilities of the excellent parallel processing and global feature extraction and BiGRU abilities of the higher computing speed and bidirectional feature extraction compared with BiLSTM under the same performance to realize the named entity recognition task of Chinese forest disease texts.

The remainder of this paper is structured as follows. Section 2 introduces the relevant research. Section 3 introduces the forest disease data set, describes the embedding of multi-features and the framework of the forest disease text, named the entity recognition method based on Transformer-BiGRU-CRF, along with the existing methods. Section 4 introduces the experimental parameters, as well as an analysis of the experimental results. Section 5 discusses the experimental results. Section 6 summarizes the text.
2. Related Work

Named entity recognition (NER) was formally established as a sub-task of information extraction in the Sixth Message Understanding Conference (MUC-6) [8], which stipulates that named entities include personal names, place names, and organization names. In the subsequent MET-2 [9] of MUC-7 and a series of international conferences, including IEER-99, CoNLL-2002, CoNLL-2003, IREX, and LREC, named entity recognition was regarded as a designated task in the field of information extraction. Moreover, the goals related to named entities are expanding.

There are three basic kinds of named entity recognition methods: rule-based methods, statistical machine learning methods, and deep learning methods. The methods based on rules rely on the manual construction of dictionaries and knowledge bases, and mostly adopt rules manually constructed by language experts. The selected features include direction words, punctuation, statistical information, and other methods to match patterns with strings. The portability of these methods is poor, as they often depend on specific fields and text features. Statistical machine learning methods typically use a manually labeled corpus for training. For new fields in this method, only a small amount of modification is needed for training. Typical machine learning models include maximum entropy (ME) [10], support vector machine (SVM) [11], hidden Markov model (HMM) [12], and conditional random fields (CRF) [13]. In recent years, named entity recognition methods based on deep learning have become the mainstream. Deep learning models are end-to-end models [14]. Deep neural networks can carry out non-linear transformations on data, then automatically learn more complex features to complete the training and prediction tasks of multi-layer neural networks. Collebert et al. [15] first proposed a named entity recognition method based on neural networks. This method limits the use of context to a fixed window size around each word, abandons the useful long-distance relationships between words, and cannot resolve the problem of long-distance dependence. With the progress of recurrent neural networks (RNNs), in terms of structure and rapid development of hardware performance, the training efficiency of deep learning has made great breakthroughs. The use of recurrent neural networks has become increasingly common. Variants of cyclic neural networks, long short-term memory (LSTM), and gated recurrent units (GRUs) have made breakthroughs in the field of natural language processing (NLP). LSTM has a strong ability to extract long-term sequence features. Huang et al. [16] first applied the bidirectional LSTM-CRF model to the benchmark sequence-marked data set of natural language processing. Bidirectional LSTM can preserve long-term memory and make use of past and future sequence information. Moreover, this model adds CRF as a decoding tool. Their experimental results showed that this model has less dependence on the word embedding and achieved a good training effect. Yang et al. [17] proposed a deep-seated recurrent neural network for sequence labeling, which uses GRUs to encode morphological and contextual information at the character and word levels, and applies a CRF field layer to predict labels. GRUs have higher calculation speeds, as they simplify the gating unit on the basis of similar accuracy to that of LSTM. Their model obtained a 91.20% F1 value on the CoNLL2003 English data set, and effectively solved the problem of cross-language joint training. The transformer model was proposed by Vaswani et al. [18], which constructs the coding layer and decoding layer through the use of a multi-head attention mechanism. Through parameter matrix mapping, the attention operation is carried out. Then, the process is repeated many times. Finally, the results are spliced to obtain the global features. As the transformer model has the advantages of parallel computing and deep architecture, it has been widely used in named entity recognition tasks. However, the transformer model does not incorporate information related to location relationships. Therefore, Yan et al. [19] improved the model to solve the problem of the transformer not being able to capture the direction information and relative position, and proposed a transformer encoder for the NER (TENER) model, which includes an attention mechanism which simultaneously captures the corresponding position and direction information. This model was implemented in the MSRA Chinese corpus, the English OntoNotes5 0 data set, and other data sets.
it was shown to be better than the original transformer model. The recognition task of nested named entities has always been the research difficulty of named entity recognition in various languages. A nested named entity is a special form of named entity, which has a complex hierarchical structure, so it is difficult to accurately identify the type of entity. For the problem of nested named entity recognition, Ankit Agrawal et al. [20] conducted in-depth research and proposed a method based on Bert to solve the problem of nested named entity recognition, which achieved the best experimental results in multiple data sets. The experiments show that the proposed method based on Bert is a more general method to solve the problem of nested named entities compared with the existing methods. With the proposal of multilingual information extraction tasks, research on multilingual named entities began to increase. In the task of Chinese named entity recognition, due to the complex properties of Chinese named entities, such as a lack of word boundaries, uncertain length, and the rich semantics of a single word, Chinese named entity recognition is more difficult than English named entity recognition. Researchers have carried out significant exploratory research on Chinese named entity recognition in different fields. Dong et al. [21] first applied the character-level BiLSTM-CRF model to the task of Chinese named entity recognition and proposed the use of Chinese radicals as the feature representation of the character, which achieved good performance without the use of Chinese word segmentation. This result indicated that Chinese named entity recognition based on a single character can achieve good results. Xuan Z et al. [22] proposed a film critic name recognition method based on multi-feature extraction. This method uses corpus to extract character features, and uses the BiLSTM-CRF model for sequence annotation. This method can adequately solve the problems of complex appellations and unlisted words in Chinese film reviews. Li Dongmei et al. [7] proposed a BCC-P named entity recognition method for plant attribute texts based on BiLSTM, CNN, and CRF. The CNN model was used to further extract sentence depth features. The accuracy reached 91.8%. The deep learning model was used to solve the problem of named entity recognition in plant attribute texts. Li Bo et al. [23] proposed a neural network model based on the attention mechanism using the Transformer-CRF model in order to solve the problem of named entity recognition for Chinese electronic cases, and achieved a 95.02% F1-value in the constructed corpus set, with better recognition performance.

By comprehensively comparing the above named entity recognition models, in this paper, we enhance the accuracy of the model by integrating character radicals, word boundaries, and part-of-speech features and also incorporate the relative position information, in order to improve the inherent problem of the transformer model being unable to capture the position information, and the BiGRU model to extract the deep features of the sentence to obtain the optimal labeling of disease names, damage sites, and pharmaceutical entities.

3. Materials and Methods

3.1. Construction and Analysis of Data Set

The text data sources of forest diseases used in this paper were mainly obtained from the Forestry Disease Database of the China Forestry Information Network [24] and Forest Pathology [25]. The Forestry Disease Database in the Forestry Information Network provides semi-structured table data of common diseases of various tree species. We used the rule-based method to extract the text information of control measures of forest diseases in a semi-structured table, including effective information such as forest disease names, control agents, and damage sites. Forest Pathology includes symptom characteristics and control measures of forest diseases, and selects the description documents related to attributes. We took the relevant information set of control measures of forest diseases in the two data sources as the final data set, with a total of 8346 relevant documents.

The extracted data were further processed as follows. Firstly, we removed invalid symbols, such as HTML symbols and meaningless punctuation. Then, we segmented overly long sentences and spliced overly short sentences. Additionally, we replaced numbers in order to improve the generalization ability of the model. According to the existing agent ontology
and forest disease ontology, the data set was uniformly labeled using the BIO labeling system. In the BIO annotation system, the B-prefix represents the first word of the entity, the I-prefix represents a word in the middle of the entity, and O represents other irrelevant words. After labeling, we wrote an automatic error correction program to detect the labeling effect. Then, we extracted and labeled the radical, part of speech, and word boundary information of each character, and obtained the final experimental self-built data set. Of these data, 80% were used as the training set and 20% as the test set. The entity categories and the distributions of the training and test sets are shown in Tables 1 and 2, respectively.

Table 1. Types of named entities.

<table>
<thead>
<tr>
<th>Types</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forest disease entities (D)</td>
<td>B-D, I-D</td>
</tr>
<tr>
<td>Drug entities (T)</td>
<td>B-T, I-T</td>
</tr>
<tr>
<td>Damage site entities (L)</td>
<td>B-L, I-L</td>
</tr>
</tbody>
</table>

Table 2. Distribution of the data sets for named entities.

<table>
<thead>
<tr>
<th>Types</th>
<th>Total Quantity</th>
<th>Training Set</th>
<th>Testing Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>5468</td>
<td>4803</td>
<td>665</td>
</tr>
<tr>
<td>T</td>
<td>7169</td>
<td>6285</td>
<td>884</td>
</tr>
<tr>
<td>L</td>
<td>706</td>
<td>25</td>
<td>81</td>
</tr>
</tbody>
</table>

3.2. The Proposed Approach

The architecture of the named entity recognition method for forest disease texts based on multi-feature fusion and Transformer-BiGRU-CRF proposed in this paper is shown in Figure 1, which is composed of four parts: a multi-feature embedding layer, a transformer layer, a BiGRU layer, and a CRF layer. By constructing word vector tables, radical vector tables, word boundary vector tables, and part-of-speech tables, and then splicing them, the model obtains a distributed vector representation of sentences as the input to the transformer layer. Then, the transformer module models the context distance and learns the implicit feature representation of the sentence, which is input to the BiGRU layer. BiGRU is used to extract the deep features of sentences. Finally, through the CRF layer’s learning constraints, the optimal global sequence label is obtained.

Figure 1. Named entity recognition method for forest disease texts.
3.2.1. Multi-Feature Embedding Layer

As the entities in forest disease texts have the characteristics of nested relationships, a large number of rarely used Chinese characters, and a lack of labels, we selected word vectors as the input for the named entity recognition model. Common sentence vector representation methods include one-hot coding and Word2Vec. The vector sparsity of one-hot coding is very high, while the Word2Vec model can use continuous and dense vectors to describe the characteristics of words. Therefore, the Word2Vec model was selected as the word-embedding model.

The named entity recognition task should fully discover and make use of the context and internal features of named entities. We combined two granularity features, at the characteristic level and word level, to further improve the recognition ability of the model. Through feature analysis of forest disease texts, the following features were selected:

1. The radical features of Chinese characters. By analyzing texts of forest diseases, it was found that many diseases and fungicides have specific radicals. For example, the names of the diseases contain partial radicals such as “口”, “疒”, and “艹”, while the control agents typically contain partial radicals such as “氵”, “雨”, and “刂”. Thus, the radical of a character was regarded as a basic feature set.

2. Word boundary features. The place names and organization names contained in the general domain data set contained obvious word boundaries. For example, most of the place names contained obvious boundary words such as “省” and “市”. In the texts on forest diseases, a large number of entities do not have obvious boundary characteristics, such as “混灭威” and “百菌清”. Therefore, the word boundary was introduced as a feature. The sentence is automatically labeled with a word boundary.

3. Part-of-speech features. Parts of speech contain the deep information of words, which is a common feature in Chinese natural language processing. By analyzing forest disease texts, it is found that there are certain rules in the part-of-speech distribution of some entities; for example, some disease entities are connected by multiple nouns, while control agent entities usually appear after verbs. This we the result of automatic part-of-speech tagging as a basic feature. Parts of speech include more than 30 kinds of nouns, verbs, prepositions, and adverbs.

By extracting word boundary features, partial radical features, and part-of-speech features, three kinds of artificial features and word vector splicing were used as the input for the transformer layer.

3.2.2. Transformer Encoder Layer with Position Information

The transformer model is completely based on an attention mechanism. Instead of using the original cycle and convolution structure, it adopts the self-attention mechanism, which can outperform RNNs and CNNs in tasks such as machine translation. The transformer model processes the characters in the sequence in parallel, and uses the self-attention mechanism to directly model the relationship between the characters in the sentence. It has better computational performance. The transformer model includes two main components: an encoder and a decoder. Because the decoder is often used for the generation task, the model proposed in this paper only uses the transformer encoder to model the context distance features. The specific structure is shown in Figure 2.

First, the input text sequence is embedded with multiple features to obtain the multi-feature embedded input sequence \( x \in \mathbb{R}^{K_x} \), where \( K_x \) represents the size of the input batch in the sequence thesaurus. As the original transformer model does not have the ability to capture sequential sequences, the positional encoding feature is added to represent the absolute position information of each character. Then, we calculate the sine transform, according to Formula (1), and the cosine transform, according to Formula (2), in order to encode the position and obtain the relative position information.

\[
PE_{(\text{pos},2i)} = \sin(\text{pos}/10,000^{2i/d_{\text{model}}})
\]

(1)
where \( pos \) refers to the position of the current character in the sentence and \( i \) refers to the dimension of the vector. In even positions, sinusoidal coding is used; in odd positions, cosine coding is used. Then, the position coding and multi-feature embedded input sequence are concatenated. The dimension is the same as that of the multi-feature input sequence, which is obtained as the input of the transformer encoder. Then, the multi-head attention mechanism layer is input and decomposed according to Equation (3). The number of heads is 8.

\[
Q_i = XW_i^Q, \quad K_i = XW_i^K, \quad V_i = XW_i^V, \quad i = 1, \ldots, 8
\]

where \( X \in \mathbb{R}^{n \times d_k}, W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d_k \times d_{model}} \), and \( d_k \) is the dimension of the multi-feature embedding space. Then, we calculate the self-attention according to Equation (4) and learn a weight for each input word. The multi-head attention mechanism is equivalent to the integration of the different self-attention mechanisms of the head. The weighted characteristic matrices are spliced and calculated according to Equation (5), in order to obtain a large characteristic matrix.

\[
Z_i = \text{Attention}(Q_i, K_i, V_i)_i = \text{softmax}\left(\frac{Q_iK_i^T}{\sqrt{d_k}}\right)V_i
\]

\[
Z = \text{MultiHead}(Q, K, V) = \text{Concat}(Z_1, \ldots, Z_8)W^O
\]

\[
\text{FFN}(Z) = \max(0, ZW_1 + b_1)W_2 + b_2
\]

where \( Z_i \in \mathbb{R}^{n \times d_{model}} \). Then, we calculate the residual connection to solve the problem of gradient disappearance when the number of back-propagation layers is large, such that the gradient can reach the input layer quickly. Then, we normalize the data (i.e., normalizing the output to the standard normal distribution), accelerate the convergence, and increase the training speed. Finally, full connection of the two layers is carried out, according to Equation (6), and the dimension transformation is completed, which enhances the expression ability of the model. In the other layers, the input is the output of the previous layer.

![Figure 2. Structure of the transformer encoder.](image-url)
3.2.3. BiGRU Layer

The GRU model is based on the LSTM model. By improving the structure of the input gate, forget gate, and output gate of the LSTM model, and combining the forget gate and input gate into the update gate, the GRU model mixes the cell and hidden states and uses the reset and update gates to jointly control the cell state of the storage of historical and future information. Due to the relatively simple structure and fewer parameters, GRU has faster calculation speed and better generalization ability. This layer is composed of different directions. The context information is extracted from the two directions, and then combined to extract the comprehensive features. The calculation process is shown in Equations (7)–(10):

\[ z_t = \sigma(w_z \cdot [h_{t-1}, C_t]) \]  
\[ r_t = \sigma(w_r \cdot [h_{t-1}, C_t]) \]  
\[ \tilde{h}_t = \tanh(w_t \cdot [r_t \cdot h_{t-1}, C_t]) \]  
\[ h_t = (1 - z_t) \cdot h_{t-1} + z_t \cdot \tilde{h}_t \]  

where \( w_z, w_r, w_t \) are the weight matrices of the update gate, reset gate, and candidate hidden state, respectively; \( C_t \) is the input of the unit in the current step; and \( z_t \) is the update gate. The gate reads the current input data and the previous memory information, then converts them into a value between 0 and 1 through the sigmoid function as the gating structure. The update gate is used to control the degree to which the state information at the previous step is brought into the current state. With a larger value of the update gate, more state information from the previous step is brought in. Furthermore, \( r_t \) is the reset gate, which controls how much information in the previous state is written to the current candidate set \( \tilde{h}_t \). A smaller reset gate means less information is written from the previous state. The update gate, reset gate, and hidden state are calculated by Equation (10) in order to obtain the output \( h_t \) of the current step.

3.2.4. CRF Layer

Taking the context-implicit features extracted by the transformer encoder layer and BiGRU layer as the input, the CRF layer ensures that the predicted tags comply with the rules through the learned constraints; for example, the label of the first character of a sentence should be “B-” or “O”, not “I-”. For an output sequence of the BiGRU layer, which is also the input sequence \( X = \{x_1, x_2, \ldots, x_n\} \) of the CRF layer, the corresponding output tag sequence is \( y = \{y_1, y_2, \ldots, y_n\} \) (\( n \) is the dimension of the input sequence). The evaluation score of the CRF definition is shown in Equation (11):

\[ s(X, y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{y_i} \]  
\[ p(y|X) = \frac{e^{s(X, y)}}{\sum_{\hat{y} \in Y_x} s(X, \hat{y})} \]  

where \( A_{y_i, y_{i+1}} \) represents the probability of transferring from the tag \( y_i \) of the position \( i \) in the annotation sequence to the tag \( y_{i+1} \) at position \( i+1 \). Additionally, \( P_{y_i} \) represents the probability that the mark is tagged \( y_i \) at position \( i \) of the sequence. The probability of calculating the tag sequence \( y \) is shown in Equation (12), in which \( \hat{y} \) represents the real tag, that is, the maximizer of \( p(y|X) \).

3.3. The Existing Methods

3.3.1. BiGRU-CRF Method

The BiGRU-CRF model is a combination of the BiGRU layer and the CRF layer. As described above, GRU controls the cell state of storing historical and future information.
through two gating structures: update gate and reset gate. The one-way GRU layer can only obtain sequence information from one direction, but BiGRU can obtain context relationship features from the forward and reverse directions. The CRF layer ensures that the predicted tags comply with the rules through the learned constraints. The structure of the model is shown in Figure 3. Firstly, the text sequences are obtained through pretraining as the input of the BiGRU layer. Secondly, the BiGRU layer is used for context modeling and feature extraction. Finally, the CRF layer is used to decode and obtain the global optimal annotation.

**Figure 3.** The structure of BiGRU-CRF method.

### 3.3.2. BiLSTM-CRF Method

The BiLSTM-CRF model is a combination of the BiLSTM layer and the CRF layer. LSTM controls the transmission and storage of information through three gating structures: input gate, forgetting gate, and output gate. Bidirectional BiLSTM is an extension of LSTM. It obtains the characteristics of context relationship from the forward and reverse directions, and considers the context-dependent information.

The structure of the model is shown in Figure 4. Firstly, the text sequences are obtained through pretraining as the input of the BiLSTM layer. Secondly, BiLSTM is used for context modeling and feature extraction. Finally, the CRF layer is used to decode and obtain the global optimal annotation.

**Figure 4.** The structure of BiLSTM-CRF method.

### 3.3.3. Transformer-BiLSTM-CRF Method

Transformer-BiLSTM-CRF is a combination of a transformer layer, a BiLSTM layer, and a CRF layer. The structure of this method is shown in Figure 5. Firstly, the text sequences are obtained through pretraining as the input of the transformer layer. Then, the transformer module models the context distance and learns the implicit feature representation of the
sentence, which is input to the BiLSTM layer. BiLSTM is used to extract the deep features of sentences. Finally, through the CRF layer’s learning constraints, the global optimal annotation is obtained.

3.3.2. BiLSTM-CRF Method

The BiLSTM-CRF model is a combination of the BiLSTM layer and the CRF layer. LSTM controls the transmission and storage of information through three gating structures: input gate, forgetting gate, and output gate. Bidirectional BiLSTM is an extension of LSTM. It obtains the characteristics of context relationship from the forward and reverse directions, and considers the context-dependent information.

The structure of the model is shown in Figure 4. Firstly, the text sequences are obtained through pretraining as the input of the BiLSTM layer. Secondly, BiLSTM is used for context modeling and feature extraction. Finally, the CRF layer is used to decode and obtain the global optimal annotation.

3.3.3. Transformer-BiLSTM-CRF Method

Transformer-BiLSTM-CRF is a combination of a transformer layer, a BiLSTM layer, and a CRF layer. The structure of this method is shown in Figure 5. Firstly, the text sequences are obtained through pretraining as the input of the transformer layer. Then, the transformer module models the context distance and learns the implicit feature representation of the sentence, which is input to the BiLSTM layer. BiLSTM is used to extract the deep features of sentences. Finally, through the CRF layer’s learning constraints, the global optimal annotation is obtained.

4. Experimental Setup and Results

4.1. Experimental Parameter Setup

We used Python version 3.6.13 to code the program and modeled it based on TensorFlow version 1.14., and we used Word2Vec to generate character vectors and the dropout algorithm to prevent overfitting of the model. The experimental hyper-parameters used in the experiment are shown in Table 3.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Character embedding size</td>
<td>230</td>
</tr>
<tr>
<td>Maximum length of sequence</td>
<td>100</td>
</tr>
<tr>
<td>Learning rate</td>
<td>0.0001</td>
</tr>
<tr>
<td>Dropout rate</td>
<td>0.1</td>
</tr>
<tr>
<td>Batch size</td>
<td>128</td>
</tr>
<tr>
<td>Number of epochs</td>
<td>100</td>
</tr>
</tbody>
</table>

In order to complete multi-feature embedding, Word2Vec was used to generate character vectors. The maximum sequence length was 100. Then, the multi-features of the character were spliced. The vector dimension was 230 in total. In batch processing, 128 forest disease control statements were processed in each batch. A larger batch size will make the descending direction of the training more accurate, cause a smaller shock, and further increase the data processing speed and memory utilization. The learning rate range for the model was [0.01, 0.001, 0.0001], for which we selected 0.0001 in this work in order to maintain the stability of the model training. At the same time, the dropout rate was set to 0.1, in order to prevent overfitting of the model.

To assess the experimental results, we used the precision (P), recall (R), and F1-score (F1), which are commonly used in the field of named entity recognition as evaluation indices. The calculation method for each evaluation index is shown in Equations (13)–(15):

\[
\text{Precision} = \frac{TP}{TP + FP} \times 100\% 
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \times 100\%
\]
\[ F1 - score = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100\% \] (15)

where TP is the number of positive samples that are correctly predicted, FP is the number of positive samples that are incorrectly predicted, and FN is the number of negative samples that are incorrectly predicted.

4.2. Experimental Results

4.2.1. Results of Multi-feature Embedding

The first part of the experiment was based on the Transformer-BiGRU-CRF model proposed in this paper. Multi-features were embedded in the input layer, including char embedding, radical embedding, POS embedding, boundary embedding, and RBP embedding. We compared and analyzed the rationality of the multi-feature embedding. The corresponding named entity recognition performance is shown in Table 4.

Table 4. Model multi-feature embedding results.

<table>
<thead>
<tr>
<th>Features</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Char</td>
<td>90.70</td>
<td>89.06</td>
<td>89.87</td>
</tr>
<tr>
<td>POS</td>
<td>83.03</td>
<td>73.19</td>
<td>77.80</td>
</tr>
<tr>
<td>Boundary</td>
<td>88.40</td>
<td>82.63</td>
<td>85.42</td>
</tr>
<tr>
<td>Radical</td>
<td>91.68</td>
<td>90.66</td>
<td>91.17</td>
</tr>
<tr>
<td>RBP</td>
<td>93.16</td>
<td>92.97</td>
<td>93.07</td>
</tr>
</tbody>
</table>

It can be seen from the comparison that the precision of the RBP embedding was 93.16%, the recall rate was 92.97%, and the F1 value was 93.07%. The effect of RBP embedding was obviously better than the result without embedding. This indicates that multi-feature embedding is in line with the characteristics of forest disease texts, effectively enhancing the performance of the named entity recognition model. The accuracy was improved by 2%. In addition, in the experimental results for word boundary and part-of-speech embedding, the recall value was somewhat low, relative to the accuracy, which may be due to the inaccurate labeling of a small part of the data in the automatic labeling of word boundaries and parts of speech, thus affecting the efficiency of the model.

4.2.2. Results of the Methods in Two Different Conditions

The second experiment was carried out to verify the effectiveness of the model. We selected four groups of models—the BiLSTM-CRF model, the BiGRU-CRF model, the Transformer-BiLSTM-CRF, and the Transformer-BiGRU-CRF model for forest disease named entity recognition proposed in this paper for comparative experiments. Then, we compared the models in two environments: character vector embedding and multi-feature fusion embedding. The experimental results are shown in Tables 5 and 6.

Table 5. Results of the different methods for character vector embedding.

<table>
<thead>
<tr>
<th>Models</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>BiLSTM-CRF</td>
<td>85.20</td>
<td>86.85</td>
<td>86.02</td>
</tr>
<tr>
<td>BiGRU-CRF</td>
<td>87.42</td>
<td>86.55</td>
<td>86.98</td>
</tr>
<tr>
<td>Transformer-BiLSTM-CRF</td>
<td>88.98</td>
<td>88.35</td>
<td>88.66</td>
</tr>
<tr>
<td>Transformer-BiGRU-CRF</td>
<td>90.70</td>
<td>89.06</td>
<td>89.87</td>
</tr>
<tr>
<td>RBP-Transformer-BiGRU-CRF</td>
<td>93.16</td>
<td>92.97</td>
<td>93.07</td>
</tr>
</tbody>
</table>

Results of the different methods for character vector embedding are shown in Table 5. The Transformer-BiGRU-CRF model constructed in this paper was better than other methods, with precision of 90.11%, recall rate of 92.37%, and F1 value of 91.02%, which shows that this method can effectively model the text of forest diseases and has adequate adaptability for the forest disease texts. The precision of the BiLSTM-CRF model was 85.20%, indicating that the BiLSTM network structure can extract the implicit context features, effectively solve the sequence problem, and complete the named entity recognition task.
in forest disease text data sets. The precision of the BiGRU-CRF model was 2% higher than that of the BiLSTM-CRF model, which shows that the simplified gating structure of the BiGRU network structure has a better generalization ability and better performance in forest disease text data sets. The experimental results for the Transformer-BiLSTM-CRF model showed that the accuracy and F1-values were slightly improved compared with those of the BiLSTM-CRF model. The introduction of a transformer decoding layer model can, therefore, enhance the feature extraction ability of the model and improve the recognition efficiency.

Table 6. Results of the different methods for multi-feature embedding.

<table>
<thead>
<tr>
<th>Models</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBP-BiLSTM-CRF</td>
<td>89.79</td>
<td>85.64</td>
<td>87.67</td>
</tr>
<tr>
<td>RBP-BiGRU-CRF</td>
<td>91.74</td>
<td>89.16</td>
<td>90.43</td>
</tr>
<tr>
<td>RBP-Transformer-BiLSTM-CRF</td>
<td>90.11</td>
<td>92.37</td>
<td>91.20</td>
</tr>
<tr>
<td>RBP-Transformer-BiGRU-CRF</td>
<td>93.16</td>
<td>92.97</td>
<td>93.07</td>
</tr>
</tbody>
</table>

Results of the various models for multi-feature embedding are shown in Table 6. The comparative experiment was conducted for each model after the multi-feature embedding input. The precision and F1-values for each algorithm based on multi-feature embedding were improved. The RBP-Transformer-BiGRU-CRF model proposed in this paper has the precision of 93.16%, the recall rate of 92.97%, and the F1-value of 93.07%. Under the condition of multi-feature embedding, the model significantly outperforms other models, and obtains favorable experimental results.

5. Discussion

In this study, we examined the named entity recognition of forest disease texts. The research mainly includes two parts: multi-feature embedding and the named entity recognition method for forest disease texts based on Transformer-BiGRU-CRF.

The first part of the paper is the discussion of multi-feature embedding. We analyzed the characteristics of the texts and determined that many diseases and fungicides have specific radicals, a large number of entities do not have obvious boundary characteristics, and there are certain rules in the part-of-speech distribution of some entities. Three features, namely, word boundary features, partial radical features, and part-of-speech features, were thus selected for multi-feature embedding experiment verification.

The second part of the paper is the discussion of the named entity recognition method for forest disease texts based on Transformer-BiGRU-CRF. We found that the RBP-Transformer-BiGRU-CRF model proposed in this paper outperformed the other mainstream algorithms under two conditions of multi-feature embedding and only character vector embedding. The experimental results show that this method can effectively model the text of forest diseases and has suitable adaptability for the forest disease text. The introduction of the transformer encoder can address the problem related to the long-term dependence of the model. The addition of the BiGRU network structure improved the ability of the model to supplement deep features, had a better extraction ability for hidden features, and made it able to adequately solve the task of named entity recognition in forest disease texts.

There are numerous domain-specific entity concepts and rare words in the collected forest disease texts. RBP-Transformer-BiGRU-CRF not only takes into account the text features of forest diseases, but also makes full use of transformer abilities of the excellent parallel processing and global feature extraction and BiGRU abilities of the higher computing speed and bidirectional feature extraction compared with BiLSTM under the same performance. This model has achieved satisfactory results in the named entity recognition task of Chinese forest disease texts.
6. Conclusions

There are many named entities in texts concerning forest diseases. However, there is also a lack of large-scale annotated data sets in this field. This study was committed to the research of named entity recognition in the field of forest diseases. For this purpose, we constructed a text entity recognition data set for forest diseases. According to the text features in this field, a named entity recognition method for forest disease texts based on the Transformer-BiGRU-CRF model was proposed. Various features, such as character radicals, word boundaries, and parts of speech, were introduced to improve the recognition ability of the model. This method fully considers the features of the text data set and the implicit features in the sentences in order to complete the optimal annotation of forest disease text sequences. Moreover, the proposed method was compared with mainstream named entity recognition methods, and an experimental comparison for each method was carried out under the condition of multi-feature embedding. The experimental results showed that the introduction of multi-features can effectively improve the accuracy of model recognition. Additionally, the method proposed in this paper obtained favorable precision, recall, and F1-values. However, the method also has shortcomings. There is no in-depth research on the phenomenon of nested named entities. The next step in nested named entity recognition can be carried out after constructing a nested entity experimental data set.

Author Contributions: Conceptualization, Q.W.; methodology, Q.W.; software, Q.W.; validation, Q.W.; formal analysis, Q.W.; investigation, Q.W.; writing—original draft preparation, Q.W.; writing—review and editing, Q.W. and X.S.; visualization, Q.W.; supervision, X.S.; project administration, X.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (no. 41971397).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References


