ABSTRACT  High quality agricultural named entity recognition (NER) model can provide effective support for agricultural information extraction, semantic retrieval and other tasks. However, the existing models ignore the potential characteristics of Chinese characters, resulting in the lack of internal semantics. Moreover, the agricultural text sequence is long, which leads to the lack of long-distance dependence of model capture. In order to solve the above problems, a self-attention mechanism RSA-CANER agricultural named entity recognition model is proposed which incorporating the potential characteristics of Chinese characters. First, the model takes character features and potential features of Chinese characters as input to enrich semantic information. Among them, character features are obtained based on ALBERT pre training tool, radical features are extracted based on convolutional neural network (CNN), and stroke features are extracted based on bidirectional long short-term memory model (BiLSTM). Then, based on the BiLSTM, the sequence characteristic matrix is obtained, and the self-attention mechanism is used to further enhance the ability of the model to capture long-distance dependence. Finally, the global optimal sequence is generated based on conditional random field (CRF) model. It obtains an F-score of 95.56%. The experimental results show that the model learns semantic information at multiple fine-grained levels of radicals and strokes, enriches the vector expression of target words, and its recognition precision is better than other models, improving the generalization ability of the model.

INDEX TERMS  Agriculture, named entity recognition, self-attention, long short-term memory, conditional random field.

I. INTRODUCTION

With the rapid development of agricultural information technology, farmers conduct online consultation through the agricultural technology service platform. In the face of massive question and answer data, how to quickly and accurately locate keywords and mine deep semantic relationships is an urgent problem to be solved by the agricultural intelligent question and answer system [1], [2]. The main task of agricultural named entity recognition is to identify different types of entities from unstructured question and answer data, such as crop diseases and pests, crop varieties, pesticide names, etc. It is the key technical link to build an intelligent question and answer system, and finally provide professional and personalized decision-making information services for grassroots agricultural technicians ([3], [4], [5]).

The research on named entity recognition in agricultural field started late, and there is a lack of standardized and open corpus. Moreover, the diversity of agricultural entities and the existence of a large number of nested entities, abbreviations and alias entities directly hinder the further development of tasks related to natural language processing in the agricultural field.

In NER task, entity recognition is treated as a sequence annotation task based on statistical machine learning [6], [7]. Common models include hidden Markov model [8], maximum entropy model [9] and conditional random field [10]. In the field of agriculture, Li et al. [11] proposed a named entity recognition method based on conditional random fields. Through feature combination and adjusting the size of context window, the precision of identifying named entities...
of crops, pests and pesticides is 97.72%, 87.63% and 98.05% respectively. Malarkodi et al. [12] designed a data set in the agricultural field containing 19 types of labels, and carried out experiments with different feature combinations. The overall precision of the system reached 83.24%. Although, the method based on machine learning shows good performance. However, this method relies on manually designed feature templates, which not only improves the performance of the model, but also reduces the robustness and generalization ability of the whole model [13], [14].

With the rise of deep learning technology, relying on its strong feature representation ability, researchers based on deep learning methods, constantly optimize and update the model framework, and have achieved good results in NER tasks ([15], [16], [17]), the existing literature research results are shown in Table 1. Researchers use the word2vec [18] tool to pre-train to obtain word vectors as input to the model. Zhang et al. [19] used word2vec to train word embedding features and character embedding features, combined dictionaries and similar words into the BiLSTM-CRF model, and built the MIFM model. However, the word vector generated by word2vec is static and has a single representation, which cannot solve the problem of polysemy. In order to better extract text feature information and solve the problem of polysemy, BERT [20], [21] is widely used in NER tasks. Although BERT model has strong learning ability, the training of dynamic word vector model takes a long time and the resource cost is high. In order to speed up the training time and consider the performance of the model, this paper introduces the lightweight BERT model ALBERT [22], which saves the model parameter space by sharing parameters.

### Table 1. Comparison of existing literature research results.

| Research | Model | Pre-training | Attention | Additional features |
|----------|-------|--------------|-----------|---------------------|
| Zhang[19]| MIFM  | word2vec     | null      | null radical feature=POS feature |
| Wu[30]   | Att-BiLSTM-CRF | word2vec | self-attention |  |
| Zheng[33]| Att-CNN-BiLSTM-CRF | BERT | multi-headed attention |  |
| Yin [36] | CCNER | word2vec | self-attention | radical feature glyph features |
| Guo[37]  | CG-ANER | BERT | null |  |

Many researchers use BiLSTM-CRF as the benchmark model to carry out research on named entity recognition ([23], [24], [25]). BiLSTM model can capture global context information, but it cannot capture the internal correlation of entities, ignoring the local context information of text sequences [26]. Moreover, the long sequence of agricultural text challenges the ability of the model to capture long-distance dependent information. As attention mechanism has been widely used in the field of natural language processing ([27], [28], [29]), more and more researchers use attention mechanism to make up for the shortcomings of BiLSTM and improve the attention of the model to key nodes.

Wu et al. [30] proposed Att-BiLSTM-CRF model based on self-attention mechanism. Based on the self-attention mechanism, establish a direct connection between each role to learn long-term dependence. The model achieved good results in CCKS-2017 shared task 2 data set. Wei et al. [31] proposed an attention mechanism to improve the vector representation in BiLSTM, and designed different attention weight redistribution methods and fused them. The F1-score of the model on JNLPBA corpus is 73.50%. Jin et al. [32] proposed a new character based gated convolutional recursive neural network GCRA. An additional gated self-attention mechanism is used to capture global dependencies from different subspaces and any adjacent characters. Zheng et al. [33] proposed a new model Att-CNN-BiLSTM-CRF. Convolutional attention layer combines local attention mechanism and CNN to capture the relationship between local context. The global multi-head attention layer optimizes the processing of sentence level information. The recall rate of the model is 88.16%, and the precision rate is 89.33%. In order to better extract the characteristics of agricultural texts from multiple perspectives and multi-level perspectives, this paper uses the self-attention mechanism to adjust the weight of the output matrix of the BiLSTM model, obtain more abundant correlation information, and improve the attention of the model to key nodes.

The model based on attention mechanism can better capture the internal correlation of data or features, obtain rich context information, and improve the performance of the model. However, the model ignores the semantic information of Chinese characters. Chinese characters are hieroglyphics with compact structure and rich internal characteristics. Compared with English word embedding vectors, Chinese radical, stroke, pinyin and so on contain a lot of valuable semantic information and morphological information [34], [35]. Yin et al. [36] proposed a BiLSTM-CRF model based on radial level feature and self-attention mechanism. CNN is used to extract root level features and capture the internal and internal correlation of characters. The F1-score of the model on CCKS-2017 and TP-CNER data sets is 93.00% and 86.34% respectively. Guo et al. [37] designed a new framework based on three-dimensional convolutional neural network to capture the context glyph feature of each character from the perspective of image.Cao et al. [38] proposed a new Chinese character embedding learning method cw2vec. Use stroke n-gram to capture the semantic and word formation information of Chinese words, and capture richer semantic information.

However, the uniqueness of Chinese characters cannot be guaranteed by using stroke features or radical features alone. For example, the word “水稻白叶枯病 (rice bacterial blight)” and the word “大灰象甲 (big gray weevil)” have stroke numbers of “25112”, but the two words have different semantic expressions. However, words with the
same stroke number have different radical characteristics. Therefore, this paper guarantees the uniqueness of Chinese characters through the splicing of radical features and stroke features.

Aiming at the shortcomings of time-consuming and high resource cost of dynamic word vector training, in order to improve the ability of the model to capture potentially important information, an RSA-CANER named entity recognition method is proposed. Based on ALBERT model, agricultural text sequence features are captured quickly and effectively, and character level feature information is obtained; Design an embedding method of extracting stroke sequence features based on BiLSTM model and radical features based on CNN model for agricultural text, and input multi granularity feature information from character level features, radical features and stroke features to enrich text semantic information. The multi-head attention mechanism is used to redistribute the weight of vector representation in the hidden state of BiLSTM layer, and pay attention to the important areas of text sequence.

Our model can effectively improve the precision of identifying entities in the agricultural field without relying on any feature engineering. The rest of the paper is organized as follows. In Section 2, we will review our approach. Section 3 reports on the experimental setup. The fourth section analyzes and discusses the experimental results in detail. Section 5 draws conclusions.

II. RSA-CANER MODEL
This section proposes a multi feature fusion RSA-CANER agricultural named entity recognition model, which is mainly composed of four parts: embedding layer, BiLSTM layer, attention layer and CRF layer, as shown in Figure 1. The model takes ALBERT model as the pre training model to process the agricultural text corpus, and sets different labels for different positions in the text sentence to obtain the character level feature vector \( E^C \); extract the character radical feature \( E^R \) through CNN layer; stroke features \( E^S \) are extracted through the BiLSTM layer, and then multi-dimensional feature information is aggregated \( E_i \), \( E_i = E_i^C + E_i^R + E_i^S \). Then \( E_i \) input the BiLSTM model to obtain the output matrix \( H \); Based on the multi-head attention mechanism, the weight of the feature matrix \( H \) is adjusted to capture the internal dependencies between characters in the sequence. In order to measure the importance of sentences in the whole text, CRF layer is added to jointly model the output of the whole sentence to obtain the global optimal tag sequence.

A. EMBEDDING LAYER
The embedding layer is a dense vector representation or distributed representation that converts Chinese text sequences into characters or words. The input sequence \( X = \{x_1, x_2, x_3, \cdots, x_n\} \) with the length of \( n \) is defined to get the corresponding embedding matrix \( E = \{e_1, e_2, e_3, \cdots, e_n\} \) after passing through the embedding layer. The corresponding embedding feature is composed of three parts: character feature \( e_1^C \), radical feature \( e_1^R \) and stroke feature \( e_1^S \), that is \( e_1 = e_1^C \oplus e_1^R \oplus e_1^S \), \( \oplus \) representing splicing.

1) CHARACTER EMBEDDING
In agricultural texts, entities have different meanings in different contexts, and there is polysemy. For example, “油葫芦 (gryllus testaceus)” belongs to insect pest in different contexts, which endangers cotton, peanuts and other crops. It can also belong to oil hyacinth, a plant belonging to the sandalwood genus in the sandalwood family. In order to make full use of sentence context information and obtain rich character level semantic representation, this paper introduces ALBERT pre training model to complete the character level feature vector representation of corpus set.

ALBERT shares the parameters of each layer of the transformer encoder, so that the superposition of multiple layers of attention becomes the superposition of one layer of attention with the same parameters, and the amount of parameters can be greatly reduced. At the same time, the stability of the model has also been improved. For input sequences \( X = \{x_1, x_2, x_3, \cdots, x_n\} \), character level semantic information \( E^C = \{e_1^C, e_2^C, e_3^C, \cdots, e_n^C\} \) is obtained through ALBERT model.

2) RADICAL FEATURE
In crop pest entities, such as “二化螟 (chilo suppressionis)”, “黑尾叶蝉 (nephotetix cincticeps)” and “螟 (chilo)” and “鳲 (cicada)” are the radicals of “虫 (insect)”, “虫 (insect)” represents pest, which is common in pest entities; Disease entities, such as “水稻叶黑粉病 (rice leaf smut)”, “水稻霜霉病 (rice downy mildew)”, “病 (disease)” represents diseases, which are common in disease entities. Therefore, the similarity of agricultural entities can be measured through the embedding of radical features. The design of this section is based on the Radical-CNN radical extraction architecture, as shown in Figure 2. The radical sequence of “螟 (chilo)” is “虫 (insect)”, “米 (mi)”, “粤 (yue)” and “五 (six)”. Define a radical sequence \( R = (r_1, r_2, r_3 \cdots, r_m) \)
containing \( m \) radical, and the calculation formula of radical embedding feature is:

\[
E^R = \max \_\text{pool}(\text{Conv}(R))
\]  

(1)

3) STROKE FEATURE

Compared with English entity recognition sequences, Chinese character stroke sequences are similar to English character sequences. This section is designed based on the BiLSTM extraction architecture to capture the internal structural features of Chinese characters.

Strokes usually refer to the uninterrupted dots and lines of various shapes that make up Chinese characters. In the “《现代汉语通用字表》 (Modern Chinese General Character List)”, five basic strokes are specified: horizontal, vertical, apostrophe, dot and fold, which are the minimum connecting units of Chinese characters. In order to facilitate computer processing, the five types of strokes are numbered accordingly, as shown in Table 2. For example, the stroke sequence of the word “病(disease)” is “横撇捺横横折撇捺 (Right-falling/Horizontal/Left-falling/Right-falling/Horizontal/Horizontal/Vertical/Turning/Left-falling/Right-falling)”, and its corresponding coding sequence is “4134112534”.

**TABLE 2. Stroke type and number.**

| Stroke name | Horizontal | Vertical | Left-falling | Right-falling | Turning |
|-------------|------------|----------|--------------|--------------|---------|
| Type        | --         | 1        | 2            | 3            | 4       | 5       |
| ID          | 1          | 2        | 3            | 4            | 5       |

First, search the word through the dictionary to obtain its stroke sequence, assign the corresponding number, and obtain the stroke sequence of a single word. The model structure is shown in Figure 3. For a Chinese character, first extract its stroke sequence \( S = (s_1, s_2, s_3 \cdots, s_k) \) and input it into the BiLSTM model for feature extraction. Through the forward LSTM output feature sequence \( \tilde{h} \) and the reverse output sequence \( \bar{h} \), the vector of hidden layer splicing is obtained. After the \( \text{tanh} \) activation function is weighted, the final output result is obtained \( E^S \). The expression is as follows:

\[
E^S = \text{tanh} (\sum_{j=1}^{k} H_{ij})
\]

(2)

**B. BILSTM LAYER**

LSTM is a special cyclic network model, which overcomes the gradient explosion problem of RNN model in the training process. In order to accurately identify agricultural named entities, a bidirectional long short-term memory model is constructed to represent the text in two different directions: forward and reverse, so as to fully obtain the past and future feature information of the target word.

The main structure of LSTM network can be formally expressed as:

\[
i_t = \sigma(W_i h_t - 1 + U_i x_t + b_i)
\]

(3)

\[
f_t = \sigma(W_f h_t - 1 + U_f x_t + b_f)
\]

(4)

\[
\tilde{c}_t = \text{tanh}(W_c h_t - 1 + U_c x_t + b_c)
\]

(5)

\[
c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t
\]

(6)

\[
o_t = \sigma(W_o h_t - 1 + U_o x_t + b_o)
\]

(7)

\[
h_t = o_t \odot \text{tanh}(c_t)
\]

(8)

where \( \sigma \) represents \( \text{Sigmoid} \) activation function, represents Hyperbolic tangent activation function, \( i_t, f_t, o_t, c_t \) indicates the input gate, forgetting gate, output gate and memory cell at time \( T \), \( W_i, W_f, W_o, W_c \) represents Weight matrix, \( b_i, b_f, b_o, b_c \) represents Offset vector, \( \tilde{c}_t \) represents Intermediate status of input, \( x_t \) represents Input vector at time \( t \), \( h_t \) represents Output at time \( t \).

Define the model input sequence \( X = (x_1, x_2, x_3 \cdots, x_n) \) and input it to the BiLSTM layer. Through the BiLSTM network model, the text representation in the forward and reverse directions is carried out to fully obtain the past and future feature information of the target word. The vector \( X \) of the character embedding layer will be used as the input of the BiLSTM layer at time \( T \). Through the forward LSTM output feature sequence \( \tilde{h} \) and the reverse output sequence \( \bar{h} \), the vector of the hidden layer splicing will be obtained, and the final output result \( h_t \) will be obtained by weighting the \( \text{tanh} \) activation function.
C. ATTENTION LAYER
In order to further improve the performance of NER method, this section introduces self-attention mechanism to reduce the dependence on external information, capture rich dependencies from the internal level of words and characters, and further improve the performance of named entity recognition. Self-attention only completes the calculation of attention within the sequence to find the internal connection of the sequence. It is often calculated by focusing on the shrinking point product of attention. The calculation formula is:

$$Attention(Q, K, V) = \sum soft_{max}(\frac{QK^T}{\sqrt{d_k}})V \quad (9)$$

where: $Q$, $K$ and $V$ are query matrix, key matrix and value matrix respectively. First, calculate the similarity between $K$, $Q$. To ensure the stability of the gradient, divide the similarity score by $\sqrt{d_k}$, and then normalize it through $softmax$ to calculate the weight value of each node. Finally, multiply with matrix $V$ to get the output result. In order to better extract the characteristics of agricultural texts from multiple perspectives and multi-level perspectives, the weight of the output matrix of the BiLSTM model is adjusted by using the multi head attention mechanism. The calculation formula of multi head attention is:

$$head_i = Attention(QW^i_Q, KW^i_K, VW^i_V) \quad (10)$$

$$Z = Multi(head_1, head_2, \cdots , head_h)W^O \quad (11)$$

The multi head attention mechanism maps $Q$, $K$, $V$ through the parameter matrix, and then does the zoom point product attention. This process is repeated for $h$ times. Finally, the results are spliced to obtain more comprehensive feature information. $W^O_Q$, $W^O_K$, $W^O_V$ are the parameter matrix of in the $ith$ head respectively, $W^O$ is the output vector.

D. CRF LAYER
Considering the dependency between named entity tags, the model introduces CRF layer, and uses state transition matrix to predict the current tag to obtain the globally optimal tag sequence. For the output matrix $H$ of BiLSTM layer, the semantic information of key characters is extracted through the multi head attention layer, and the output result, can be expressed as the output vector corresponding to the ith word. Take the output sequence $Z$ as the input of CRF layer, and the probability of the corresponding output tag sequence is:

$$s(X, y) = \sum_{i=1}^{n} Z_{i,y} + \sum_{i=0}^{n-1} A_{y,y+1} \quad (12)$$

Then, the conditional probability of sequence $y$ is obtained by using softmax function. Finally, Viterbi algorithm is used to take the sequence with the highest score $y^*$ as the final annotation result of the model.

In the prediction process, a set of sequences that maximize the overall probability of output are:

$$y^* = \arg \max_{\hat{y} \in Y_X} S(X, \hat{y}) \quad (13)$$

III. EXPERIMENTS
A. DATA ACQUISITION
In order to solve the problem of limited corpus, collect the corresponding text data, and ensure the reliability of the data through data cleaning, denoising, de redundancy and other preprocessing. Combined with the knowledge of experts in the field, the corpus is classified and labeled, and an agricultural corpus, agent, AgNER is constructed, which contains 29483 entities, including five types of entities: crop diseases, crop pests, pesticide names, agricultural machinery names, and crop variety names. Without relying on manual design features, the recognition performance of the model is verified on the agent data set by adjusting different model parameters. The training set, test set and verification set in the corpus are allocated in the proportion of 6:2:2. There is no overlap between the data sets, so the test results of the test data set can be used as the evaluation index of entity recognition effect.

B. EXPERIMENTAL SETUP
The experimental parameters of the model are set as follows: using the ALBERT base model, there are 12 transformer layers, 768 dimensional hidden layers and 12 multi head attention mechanisms. The maximum sequence length is 256, the dimension of BiLSTM hidden layer is 128, dropout is set to 0.5, Adam optimization algorithm is used, the training learning rate is 0.001, batch processing parameter is 32, and the number of iterations is 100. The model was evaluated by three indicators: Precision, Recall and F-score.

IV. RESULT ANALYSIS
A. COMPARISON WITH DIFFERENT FEATURE
In this section, Att-ALBERT-BiLSTM-CRF is used as the benchmark model, and two additional feature information of radicals and stroke sequences are integrated into the model for comparative experiments. The results are shown in Table 3. Among them, the benchmark model Att-ALBERT-BiLSTM-CRF takes the character level feature $E^C$ as the input, and the recognition precision is 94.25%, the recall rate is 94.18%, and the F-score is 94.21%. Incorporating the radical feature $E^R$, the precision of the model is 94.71% and the F-score is 94.64%. With the stroke feature $E^S$, the precision of the model is 94.93% and the F-score is 94.99%. In order to ensure the uniqueness of Chinese characters, the $E^R + E^S$ fusion is used as an additional feature of the model and spliced with the word level feature $E^W$. The precision of the model is 95.48% and the F-score is 95.56%. The analysis shows that by integrating stroke features and radical features, the model can not only capture the stroke dependency of text words, but also enhance the semantic representation of words and improve the recognition ability of the model through the embedding of radical features.

B. COMPARISON WITH DIFFERENT CONVOLUTION KERNEL SIZES
RSA-CANER model captures the characteristics of radicals based on CNN model. In order to fully mine the context
relationship between adjacent radicals, multiple convolution networks with different window sizes are used to extract local context information of different scales. Comparative experiments are carried out to verify the influence of CNN window size on the model. The experimental results are shown in Table 4. When CNN sets the window number to 3 and the window width to 2, 3, 4, the precision of the model is 94.57% and the F-score is 94.66%. Compared with the model with 3 windows, when the number of windows is 5 and the width is 1, 2, 3, 4, 5, the precision of the model is up to 95.48%. However, when the number of windows is 5 and the width is 2, 3, 4, 5, 6, the performance of the model decreases slightly. The analysis shows that increasing the number of windows of CNN model and setting the appropriate window width can effectively increase the recognition precision of the model. With the increase of the number and width of windows, the training cost becomes higher and higher, and the performance of the model is difficult to be improved or even decreased.

TABLE 3. Performance comparison of model with different embedded vector.

| Feature | Models                        | Precision | Recall | F-score |
|---------|-------------------------------|-----------|--------|---------|
| $E^c$   | Att-ALBERT-BiLSTM-CRF         | 94.25     | 94.18  | 94.21   |
| $E^c+E^s$ | Att-ALBERT-BiLSTM-CRF     | 94.71     | 94.58  | 94.64   |
| $E^c+E^s$ | Att-ALBERT-BiLSTM-CRF         | 94.93     | 95.06  | 94.99   |
| $E^c+E^s+E^g$ | Att-ALBERT-BiLSTM-CRF | 95.48     | 95.64  | 95.56   |

D. COMPARISON RECOGNITION EFFICIENCY OF DIFFERENT MODELS

In order to verify the recognition performance of the model for five types of named entities, BiLSTM-CRF, BERT-BiLSTM-CRF, ALBERT-BiLSTM-CRF, Att-ALBERT- BiLSTM-CRF, RSA-CANER are compared. The test results are shown in Figure 4 and Figure 5. Figure 4 shows the comparison of the precision of the model, and figure 5 shows the comparison of the F-score of the model. For the BiLSTM-CRF model, the recognition precision of disease, pest, pesticide name, crop variety and agricultural machinery name entities are 92.84%, 90.64%, 86.72%, 75.74% and 75.82% respectively. Compared with the word2vec pre training tool, the ALBERT-BiLSTM-CRF model is based on the ALBERT layer to fuse the node context and learn more segment level information, location information and word level information. The model can expand the representation space of nodes, and the recognition precision of disease names, pest names and other entities with obvious boundary characteristics has reached 96.89% and 96.94%. With the introduction of attention mechanism, compared with ALBERT-BiLSTM-CRF model, the recognition precision of ATT-ALBERT-BiLSTM-CRF model in five categories of named entities has increased by 0.96, 0.31, 0.4, 2.03 and 2.24 percentage points respectively. For small-scale and complex word formation entities such as crop variety names and agricultural machinery equipment names, the model can obtain rich word vector expression through the multi head self-attention mechanism, so that the model can learn better according to the context semantic environment to improve the generalization ability of the model. The precision of the model is improved by 5.5 percentage points, and the F-score is increased by 5.25 percentage points. Compared with the BERT-BiLSTM-CRF model, ALBERT- BiLSTM-CRF completes word level embedding based on ALBERT to obtain dynamic word vectors. On the basis of ensuring the performance of the model, using the parameter sharing mechanism, the precision is improved by 0.35%. RSA-CANER model has the highest precision of 95.48%, and the F-score is 95.56%. The analysis shows that the introduction of radicals and stroke features can effectively capture the dependencies between text characters, enrich semantic information, and improve the recognition performance of the model.

TABLE 5. Performance comparison of different models.

| Models                        | Precision | Recall | F-score |
|-------------------------------|-----------|--------|---------|
| BiLSTM-CRF                    | 88.01     | 88.76  | 88.38   |
| BERT-BiLSTM-CRF               | 93.51     | 93.76  | 93.63   |
| ALBERT-BiLSTM-CRF             | 93.86     | 93.82  | 93.84   |
| Att-ALBERT-BiLSTM-CRF         | 94.25     | 94.18  | 94.21   |
| RSA-CANER                     | 95.48     | 95.64  | 95.56   |

C. COMPARISON WITH DIFFERENT MODELS

In order to verify the recognition performance of RSA-CANER model in agricultural corpus, comparative experiments were carried out with BiLSTM-CRF, BERT-BiLSTM-CRF, ALBERT-BiLSTM-CRF, Att-ALBERT-BiLSTM-CRF and other mainstream models. The experimental results are shown in Table 5. The precision of BiLSTM-CRF model is 88.01%, and the F-score is 88.38%. Compared with the BiLSTM-CRF model, the BERT-BiLSTM-CRF model introduces the BERT layer to fully extract the character level and sentence level features, and dynamically adjust the character vector representation

TABLE 4. Performance comparison of model with different convolution kernel sizes.

| Number | Width     | Precision | Recall | F-score |
|--------|-----------|-----------|--------|---------|
| 3.00   | 2,3,4     | 94.57     | 94.76  | 94.66   |
| 4.00   | 1,2,3,4   | 95.02     | 94.89  | 94.95   |
| 4.00   | 2,3,4,5   | 95.17     | 95.11  | 95.14   |
| 5.00   | 1,2,3,4,5 | 95.48     | 95.64  | 95.56   |
| 5.00   | 2,3,4,5,6 | 95.27     | 95.44  | 95.35   |
word level feature information and have better recognition precision. RSA-CANER model introduces radical and stroke features. For named entities with complex word formation, the model can capture more accurate and richer semantic information. The recognition precision of five types of named entities reaches the highest, which are 98.27%, 98.01%, 92.18%, 91.37% and 86.94% respectively, further improving the performance of the model.

V. CONCLUSION

This paper analyzes the characteristics of Chinese character word formation, constructs the radical feature and stroke sequence feature extraction model based on CNN and BiLSTM deep learning framework, obtains the potential internal features of Chinese characters in the corpus, and enriches the semantic information of the model. In order to further improve the calculation efficiency of the model, a lightweight ALBERT pre training model is introduced to obtain word level feature representation. Based on the BiLSTM-CRF model framework, a multi head self-attention mechanism is introduced to construct RSA-CANER named entity recognition method. Through the self-attention mechanism, we can obtain the internal dependencies between nodes in the sequence, better capture the semantic information, better capture the internal dependencies of the text, and extract rich semantic information. The precision of the model is 95.48%, and the F-score is 95.56%, which is suitable for NER tasks in the agricultural field.

REFERENCES

[1] J. Zhang, M. Guo, Y. Geng, M. Li, Y. Zhang, and N. Geng, “Chinese named entity recognition for apple diseases and pests based on character augmentation,” Comput. Electron. Agricult., vol. 190, Nov. 2021, Art. no. 106464.

[2] L. Li, H. Zhou, X. Guo, C. Liu, J. Su, and Z. Tang, “Named entity recognition of diseases and insect pests based on multi source information fusion,” Trans. Chin. Soc. Agricult. Machinery, vol. 52, no. 12, pp. 253–263, 2021.

[3] X. Hu, J. Duan, and D. Dang, “Natural language question answering over knowledge graph: The marriage of SPARQL query and keyword search,” Knowl. Inf. Syst., vol. 63, no. 4, pp. 819–844, Apr. 2021.

[4] P. Zhao, C. Zhao, H. Wu, and W. Wang, “Named entity recognition of Chinese agricultural text based on attention mechanism,” Trans. Chin. Soc. Agricult. Machinery, vol. 52, no. 1, pp. 185–192, 2021.

[5] B. Drury and M. Roche, “A survey of the applications of text mining for agriculture,” Comput. Electron. Agricult., vol. 163, Aug. 2019, Art. no. 104864.

[6] G. Zhou and J. Su, “Named entity recognition using an HMM-based chunk tagger,” in Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL), Philadelphia, PA, USA, 2001, pp. 473–480.

[7] J. Lei, B. Tang, X. Lu, K. Gao, M. Jiang, and H. Xu, “A comprehensive study of named entity recognition in Chinese clinical text,” J. Amer. Med. Informat. Assoc., vol. 21, no. 5, pp. 808–814, Dec. 2014.

[8] J. Zhang, D. Shen, G. Zhou, J. Su, and C.-L. Tan, “Enhancing HMM-based biomedical named entity recognition by studying special phenomena,” J. Biomed. Informat., vol. 37, no. 6, pp. 411–422, Dec. 2004.

[9] A. McCallum, D. Freitag, and F. C. Pereira, “Maximum entropy Markov models for information extraction and segmentation,” in Proc. ICML, 2000, pp. 591–598.

[10] C. Sun, Y. Guan, X. Wang, and L. Lin, “Rich features based conditional random fields for biological named entities recognition,” Comput. Biol. Med., vol. 37, no. 9, pp. 1327–1333, Sep. 2007.

[11] X. Li, X. Wei, L. Jia, X. Chen, L. Liu, and Y. Zang, “Recognition of crops, diseases and pesticides named entities in Chinese based on conditional random fields,” Trans. Chin. Soc. Agricult. Machinery, vol. 48, no. s1, pp. 178–185, 2017.

[12] S. Malarkodi, L. Elisabeth, and L. D. Sobha, “Named entity recognition for the agricultural domain,” Res. Comput. Sci., vol. 117, no. 1, pp. 121–132, 2016.

[13] F. Ye, Y. Y. Chen, G. G. Zhou, H. M. Li, and Y. Li, “Intelligent recognition of named entity in electronic medical records,” Chin. J. Biomed. Eng., vol. 30, no. 2, pp. 256–262, 2011.

[14] C. Liu, Y. Yu, X. Li, and P. Wang, “Named entity recognition in equipment support field using tri-training algorithm and text information extraction technology,” IEEE Access, vol. 9, pp. 126728–126734, 2021.

[15] J. P. C. Chiu and E. Nichols, “Named entity recognition with bidirectional LSTM-CNNs,” Trans. Assoc. Comput. Linguistics, vol. 4, pp. 357–370, Dec. 2016.

[16] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, “Neural architectures for named entity recognition,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., San Diego, CA, USA, 2016, pp. 260–270.

[17] M. Gridach, “Character-level neural network for biomedical named entity recognition.” J. Biomed. Informat., vol. 70, pp. 85–91, Jun. 2017.

[18] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in Proc. Adv. Neural Inf. Process. Syst., Lake Tahoe, NV, USA, Dec. 2013, pp. 3111–3119.

[19] N. Zhang, G. Xu, Z. Zhang, and F. Li, “MIFM: Multi-granularity information fusion model for Chinese named entity recognition,” IEEE Access, vol. 7, pp. 181648–181655, 2019.

[20] Y. Yu, Y. Wang, J. Mu, W. Li, S. Jiao, Z. Wang, P. Lv, and Y. Zhu, “Chinese mineral named entity recognition based on BERT model,” Exp. Syst. Appl., vol. 206, Nov. 2022, Art. no. 117727.

[21] Q. Zhang, Y. Sun, L. Zhang, Y. Jiao, and Y. Tian, “Named entity recognition method in health-preserving field based on BERT,” Proc. Comput. Sci., vol. 183, pp. 212–220, Jan. 2021.

[22] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “ALBERT: A lite BERT for self-supervised learning of language representations,” in Proc. 8th Int. Conf. Learn. Represent., La Jolla, CA, USA, 2020, pp. 1–17.

[23] L. Yan and S. Li, “Grape diseases and pests named entity recognition based on BiLSTM-CRF,” in Proc. IEEE 4th Adv. Inf. Manage. Communicates, Electron. Autom. Control Conf. (IMCEC), Jun. 2021, pp. 2121–2125.

[24] Y. Chen, C. Zhou, T. Li, H. Wu, X. Zhao, K. Ye, and J. Liao, “Named entity recognition from Chinese adverse drug event reports with lexical feature based BiLSTM-CRF and tri-training,” J. Biomed. Informat., vol. 96, Aug. 2019, Art. no. 103252.

[25] K. Xu, Z. Yang, P. Kang, Q. Wang, and W. Liu, “Document-level attention-based BiLSTM-CRF incorporating disease dictionary for disease named entity recognition,” Comput. Biol. Med., vol. 108, pp. 122–132, May 2019.
[26] C. Wu, G. Luo, C. Guo, Y. Ren, A. Zheng, and C. Yang, “An attention-based multi-task model for named entity recognition and intent analysis of Chinese online medical questions,” J. Biomed. Informat., vol. 108, Aug. 2020, Art. no. 103511.

[27] B. Zhang, K. Liu, H. Wang, M. Li, and J. Pan, “Chinese named-entity recognition via self-attention mechanism and position-aware influence propagation embedding,” Data Knowl. Eng., vol. 139, May 2022, Art. no. 101983.

[28] B. Tang, X. Wang, J. Yan, and Q. Chen, “Entity recognition in Chinese clinical text using attention-based CNN-LSTM-CRF,” BMC Med. Inform. Decis. Making, vol. 19, no. 3, pp. 89–97, Apr. 2019.

[29] X. Guo, S. Lu, Z. Tang, Z. Bai, L. Diao, H. Zhou, and L. Li, “CG-ANER: Enhanced contextual embeddings and glyph features-based agricultural named entity recognition,” Comput. Electron Agricult., vol. 20, no. 3, pp. 396–406, Aug. 2020, Art. no. 103511.

[30] G. Wu, G. Tang, Z. Wang, Z. Zhang, and Z. Wang, “An attention-based BiLSTM-CRF model for Chinese clinic named entity recognition,” IEEE Access, vol. 7, pp. 113942–113949, 2019.

[31] K. Zheng, L. Sun, X. Wang, S. Zhou, H. Li, S. Li, L. Zeng, and Q. Gong, “Named entity recognition in electric power metering domain based on attention mechanism,” IEEE Access, vol. 9, pp. 152564–152573, 2021.

[32] J. Qiu, Q. Wang, Y. Zhou, T. Ruan, and J. Gao, “Fast and accurate recognition of Chinese clinical named entities with residual dilated convolutions,” in Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM), Dec. 2018, pp. 935–942.

[33] M. Yin, C. Mou, K. Xiong, and J. Ren, “Chinese clinical named entity recognition with radical-level feature and self-attention mechanism,” J. Biomed. Informat., vol. 98, Oct. 2019, Art. no. 103289.

[34] X. Guo, S. Lu, Z. Tang, Z. Bai, L. Diao, H. Zhou, and L. Li, “CG-ANER: Enhanced contextual embeddings and glyph features-based agricultural named entity recognition,” Comput. Electron Agricult., vol. 194, Mar. 2022, Art. no. 106776.

[35] S. Cao, W. Lu, J. Zhou, and X. Li, “cw2vec: Learning Chinese word embeddings with stroke N-gram information,” in Proc. AAAI Conf. Artif. Intell., vol. 32, no. 1, pp. 5053–5061.

PENGFEI ZHAO was born in Henan, China. He is currently pursuing the Ph.D. degree with the College of Engineering, Shantou University, China. He is also a Research Associate with the National Engineering Research Center for Information Technology in Agriculture, Beijing. His current research interests include virtual reality, intelligent human–computer interaction devices, and their applications in the field of agricultural informatization.

MO HAN was born in Beijing, China. She received the bachelor’s degree from Southwest University, Chongqing. She is currently working with the National Engineering Research Center for Information Technology in Agriculture, and obtained the tistle of an Intermediate Engineer. Her current research interests include virtual reality, intelligent human–computer interaction devices, and their applications in the field of agricultural informatization.

WEI WANG was born in Beijing, China. He received the bachelor’s degree from Beijing Film Academy, Beijing. He is currently working with the National Engineering Research Center for Information Technology in Agriculture, and obtained the tistle of an Intermediate Engineer. His current research interests include digital art and artificial intelligence, and their applications in the field of agricultural informatization.

HAI LIU was born in Hunan, China. He received the Ph.D. degree from the China Academy of Forestry Science. He is currently working at the National Engineering Research Center for Information Technology in Agriculture, Beijing. His current research interests include data visualization, virtual reality, intelligent human–computer interaction devices, and their applications in the field of agricultural virtual interactive display.

* * *