ISSN 2616-5767 Vol.5, Issue 4: 34-42, DOI: 10.25236/AJETS.2022.050407

Research on Agricultural Named Entity Recognition Based on Pre Train BERT

Zhu Lun[1,a], Zhou Hui[1,b,*]

[1]School of Computer Science and Artificial Intelligence, Aliyun School of Big Data, Changzhou University, Changzhou, 213164, Jiangsu, China
[a]13961157353@139.com, [b]631022588@qq.com
**Corresponding author*

Abstract: *Aiming at the problems of traditional named entity recognition methods relying on artificial dictionary and insufficient feature extraction in the process of agricultural pest information extraction, considering the complexity and fuzziness of agricultural text data, an agricultural named entity recognition method based on pre training BERT is proposed. firstly, the unlabeled pre training BERT was used to eliminate ambiguity, then BILSTM was used to capture long-distance dependence, and finally the best sequence annotation was selected through CRF. In addition, combined with the text particularity of agricultural entities, the word itself and partial radicals are selected to establish a joint multi feature PBERT-BILSTM-CRF for entity recognition. experiments show that the PBERT-BILSTM-CRF combined with the combined characteristics of the word itself and the radical has improved the precision, recall and F1 value compared with other models, and its optimal F1 value has reached 90.24%. the model has the characteristics of fast training speed and strong recognition ability. Named entity recognition is the premise of many tasks in the middle and downstream of natural language processing. The model provides a research basis for the construction of knowledge graph in agricultural field and agricultural Q&A.*

Keywords: *agricultural; named entity recognition; CRF; fine tuning model; feature selection*

1. Introduction

With the emergence of big data era and the rise of intelligent agriculture, natural language processing in the field of agriculture has become more and more important[1], and agricultural named entity recognition is one of the important research directions. As an intelligent extraction method, agricultural named entity recognition mainly aims to identify proper nouns from a large number of agricultural texts, such as crop names, diseases, pests, pesticides and other entities[2]. Agricultural named entity recognition is an indispensable part of downstream tasks such as agricultural knowledge map and question answering system, and has considerable research value.

Like other named entity recognition tasks, agricultural named entity recognition is usually solved as a sequence label problem[3], in which entity boundary and category label are jointly predicted[4]. Most of the early research methods were rule-based and dictionary based. Later, traditional machine learning methods were used, such as HMM, HEMM, CRF and so on. Now, much work focuses on extracting named entities from text using depth neural network[5]. At present, the popular deep learning methods of named entity recognition are generally based on word embedding. It can learn the similar representation of words with similar semantics or functions, and then the word embedding is input into the long short term memory neural network(LSTM)and conditional random field(CRF)[6-8]. Zhang [7] et al. ignored the particularity of Chinese. The embedding of the same word in different semantic sentences is the same. However, many words have different meanings in different contexts. Another method is to add Convolutional Neural Network(CNN) based on attention mechanism to the model, for instance,Qiang[9]at al. added CNN model to medical health named entity recognition to obtain local features. However, this has little improvement in some specific fields. Another problem with word embedding methods is that training such models usually requires a large amount of tag data. In the general field, large-scale training data can often be obtained, while in the agricultural field, the labeled data is difficult to collect [3].

In recent years, many people use unsupervised pre training language models on large unmarked corpora. Radford at al.[10-11] used a two-stage approach to solve natural language processing tasks. The first stage is to train the language model on a large corpus, and the second stage is to apply the pre trained language model to downstream tasks. Due to the success of these models in various NLP tasks, especially

ISSN 2616-5767 Vol.5, Issue 4: 34-42, DOI: 10.25236/AJETS.2022.050407

when specific annotations are difficult to obtain, the use of unsupervised pre training becomes very useful. Devlin at al.[12] proposed a transformer based bidirectional encoder representation (BERT) model and made improvements in many tasks. However, for general agricultural corpus, there is no public pre training BERT language model [13]. Through the pre training of agricultural corpus and web crawler, this paper obtains a large number of agricultural texts. Our benchmark model is a publicly available BERT pre trained in general fields(https://github.com/google-research/BERT).The experimental results show that the fine-tuning BERT model pre trained on agricultural corpus has better performance than the original BERT model. The model and other models are verified on the self-built agricultural field data set, and the results show that the effect of this model is better than other models. In recent years, the radical features of Chinese characters have been widely used to enhance different Chinese natural language processing tasks [14-16]. This paper establishes a joint multi feature PBERT-BILSTM-CRF model to recognize agricultural texts by analyzing the word features and partial radicals of agricultural entities.

2. Model Structure

2.1 Overall structure of model

The overall model of this paper includes three layers:

(1) BERT layer:BERT is a pre training model, which uses the superposition input of word vector, sentence vector and position vector, and uses Transformer to encode and generate feature vector representation. This paper uses unlabeled agricultural text to pre train BERT.

(2) BILSTM layer:It is used to capture long-distance dependence and extract global features based on the pre trained BERT output.

(3) CRF layer:Calculate and output the best sequence annotation.

The model structure is shown in Figure 1.

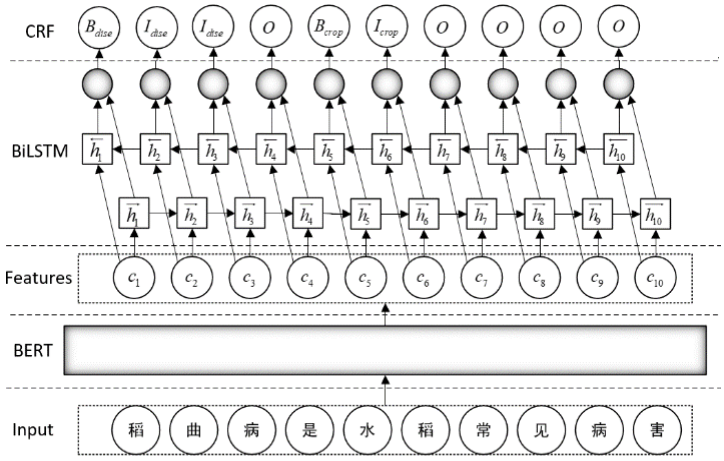

Figure 1: Overall structure diagram of model

2.2 BERT

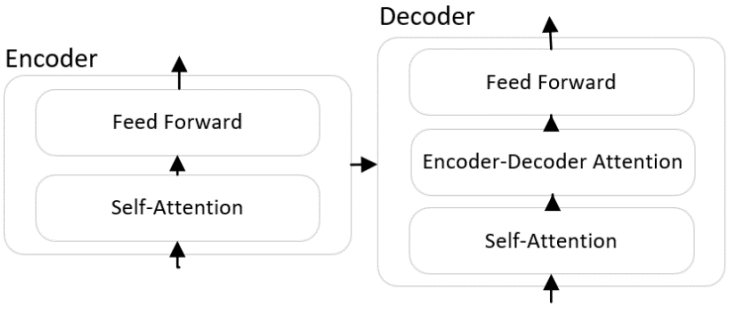

Figure 2: Encoder-Decoder architecture diagram

ISSN 2616-5767 Vol.5, Issue 4: 34-42, DOI: 10.25236/AJETS.2022.050407

BERT proposed two new unsupervised tasks, masked language model and next sentence prediction, and combines the results of the two tasks. The former is used to obtain the word level representation, and the latter is used to obtain the sentence level representation. BERT can fully learn semantic features and generate different vectors for polysemous words according to the scene. BERT can do this thanks to the self- attention mechanism in Transformer. The Transformer model adopts the encoder decoder architecture, and the model structure is shown in Figure 2.

The model will embed the input data. After embedding, it will be input to the encoder layer. After self attention processes the data, it will send the data to the feedforward neural network, and the obtained output will be input to the next encoder.

The inputs of BERT model are word vector, segment vector and position vector respectively, as shown in Figure 3. Each "E" actually input is the superposition of these three vectors, and T represents transformer. The sum of vectors is input into the Transformer network, and the output of the BERT layer is obtained by the last layer of Transformer.

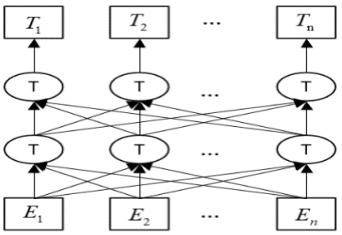

Figure 3: BERT structure diagram

The purpose of Transformer is to obtain the relationship between words, capture the internal structure of sentences, and reflect the importance and relevance of different words. The calculation formula is (1):

$$Attention(Q,K,V) = soft\max(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

In the formula(1), the three vectors Q, K and V are the result of multiplying the embedding vector by a random initialization matrix. d_K Is the dimension of the input vector [17].

2.3 BILSTM

In 1997, Hochreiter and schmidhuber proposed the LSTM model, which was originally designed to solve the gradient slowness and gradient explosion associated with recurrent neural network (RNN) training. The LSTM network structure consists of three control units called "Gates" and a memory unit. Input gate, output gate and forget gate are three gates in LSTM network structure [18].

$$\mathrm{i}_\mathrm{t} = \sigma(w_{xi}x_t + b_{ii} + w_{hi}h_{t-1} + b_{hi}) \quad (2)$$

$$f_\mathrm{t} = \sigma(w_{xf}x_t + b_{if} + w_{hf}h_{t-1} + b_{hf}) \quad (3)$$

$$o_\mathrm{t} = \sigma(w_{xo}x_t + b_{io} + w_{ho}h_{t-1} + b_{ho}) \quad (4)$$

$$\tilde{c}_\mathrm{t} = \tanh(w_{xc}x_t + b_{ic} + w_{hc}h_{t-1} + b_{hc}) \quad (5)$$

$$c_t = f_t c_{t-1} + i_t \tilde{c}_\mathrm{t} \quad (6)$$

$$h_t = o_t \otimes \tanh(c_t) \quad (7)$$

In the formula(2-7), "W" and "b" respectively represent the weight matrix and bias vector connecting the two layers, σ represents the sigmoid activation function, x_t represents the input vector, $\otimes$ represents the dot product operation, $\tilde{c}_t$ represents the state at time of "t", and h_t represents the output

ISSN 2616-5767 Vol.5, Issue 4: 34-42, DOI: 10.25236/AJETS.2022.050407

at time of "t". The key for LSTM to remember long-term dependence is input gate and forgetting gate. Its core idea is to manage the information in the storage unit by learning the parameters of the three gates in the LSTM unit, so that the useful information can be stored in the storage unit after a long time sequence.

LSTM is a one-way recurrent neural network, which can only obtain the above characteristic relationship, while the word formation of agricultural entities is complex and different. In order to achieve better recognition effect, BILSTM (Bidirectional LSTM) network model is constructed. BILSTM can effectively obtain the characteristic information of the context. The final output of BILSTM is composed of past hidden information and future hidden information[19].

2.4 CRF

CRF is a discriminant probabilistic undirected graph model, which has the ability to express the characteristics of long-distance dependence and overlap of elements [4]. CRF is widely used in the field of named entity recognition and serialization annotation. Figure 4 shows the structure of the chain conditional random field.

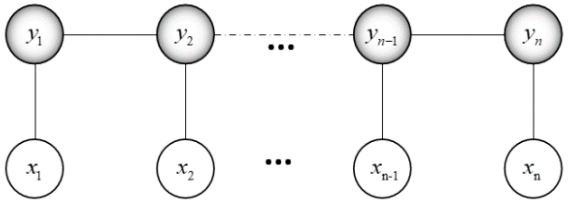

Figure 4: Chain conditional random field

Receive an input sequence $X=(x_1,x_2,...,x_n)$,The score of the prediction sequence $Y=(y_1,y_2,...,y_n)$ can be expressed as (8):

$$S(x,y)=\sum_{i=0}^{n}A_{y_i,y_{i+1}}+\sum_{i=1}^{n}P_{i,y_i} \quad (8)$$

Where "A" represents the transfer matrix and "P" is the output result of BILSTM. The softmax function is used to obtain the probability of sequence "y".

$$p(y|x)=\frac{e^{S(x,y)}}{\sum_{\tilde{y}\in Y_X}S(x,\tilde{y})} \quad (9)$$

y^* is the real tag value and Y_x is all possible tag sets. In the training process, the maximum likelihood probability of the correct tag sequence is expressed as (10):

$$\log(P(y|x))=S(x,y)-\sum_{y^*\in Y_X}^{n}S(x,\tilde{y}) \quad (10)$$

Finally, Viterbi algorithm is used to obtain the best predicted tag sequence:

$$y^*=\arg\max(S(x,\tilde{y})) \quad (11)$$

3. Data Processing

Chinese agricultural named entity recognition lacks an open corpus data set. This paper establishes an agricultural Entity Recognition Corpus through three steps: data collection, data preprocessing and data annotation. The corpus data of this paper mainly capture the text materials about crops, diseases, pests and pesticides on major agricultural websites (China agricultural information network, China agricultural knowledge network, China crop germplasm resources information network, national agricultural science data center, etc.) through the Scrapy framework. The annotated corpus contains 412

ISSN 2616-5767 Vol.5, Issue 4: 34-42, DOI: 10.25236/AJETS.2022.050407

agricultural texts, with a total of 43096 sentences. The data set is divided into training data set, verification set and test set in the ratio of 6:2:2. In order to limit the length of sentences, we separate each record with a period, and the longest sequence length is 480. The distribution of entity categories in the dataset is shown in Table 1.

Table 1: Dataset entity distribution

Entity category	Category tag	Description	Number	Proportion
Crop	Crop	Names of fruits, vegetables, cereals, grains	307	10.41%
Disease	Dise	Diseases affecting plant growth	1243	42.15%
Pest	Pest	Names of insects that damage plants	1104	37.44%
Drug	Drug	Insecticide, bactericide et al.	295	10.00%

This paper chooses to use BIO marking scheme to label the named entity, B represents the beginning of the entity, I represents the interior and end of the entity, and O represents others. In order to identify the category well, the category information is added after the entity label. The specific marking method of data is shown in Table 2.

Table 2: Example of dataset entity mark

Sentence	Mark	Sentence	Mark
稻	B-Dise	卷	B-Pest
曲	I-Dise	叶	I-Pest
病	I-Dise	螟	I-Pest
是	O	以	O
水	B-Crop	幼	O
稻	I-Crop	虫	O
常	O	为	O
见	O	害	O
病	O	大	B-Crop
害	O	豆	I-Crop

The BERT model is pre trained with agricultural text, and the recommended method in paper [12] is adopted. Starting with the checkpoint of an existing BERT, run additional pre training steps on a specific domain. According to the training process of native BERT,a .tfrecord file containing agricultural corpus text is generated. Since the original vocabulary of BERT does not contain some common characters in the field of agriculture, 34 characters are added to the vocabulary in this paper. In the experiment, the maximum sentence length is set to 480 and the probability of masking language model is set to 0.15, the maximum prediction of each sentence is 75. In addition, the maximum training step is set to100000 to train the model until the training loss stops decreasing. After the pre training process is completed, the tensor flow model ending with .ckpt is obtained. Then, we use the conversion script(https://github.com/huggingface/transformers) to convert the tensor flow model to PyTorch model ending in .bin for the next experiment.

4. Experiments and Results

4.1 Experimental environment and configuration

The experimental model is carried out on Win10, based on Python = 3.6 and PyTorch = 1.4.0, the benchmark BERT version is BERT-base-Chinese, and the GPU used is an nvida RTX 2080ti. In order to limit the length of sentences, we separate each record with a period, and the longest sequence length is 480. The model uses a bidirectional LSTM network, and the hidden layer dimension is set to 128. In order to reduce the over fitting problem of the model, the dropout mechanism is introduced. The dropout value directly affects the model performance and is set to 0.5. The backward propagation algorithm and Adam optimization algorithm are selected, the learning rate is 0.001, and the model batch_ size value is

ISSN 2616-5767 Vol.5, Issue 4: 34-42, DOI: 10.25236/AJETS.2022.050407

set to 16 and the number of iterations is set to 50. All the experimental results in this paper are obtained by taking the mean value of many experiments.

4.2 Experimental evaluation criteria

Three indexes were used to evaluate the results: Precision(P), Recall (R) and F-Measure(a=1).

The calculation formula :

$$P = \frac{TP}{TP + FP} \quad (12)$$

$$R = \frac{TP}{TP + FN} \quad (13)$$

$$F_1 = \frac{2 \times P \times R}{P + R} \quad (14)$$

Where TP is the number of correctly identified entities, FP is the number of incorrectly identified entities, and FN is the number of unrecognized entities.

4.3 Experimental comparison

4.3.1 Comparison of different models

In order to verify the performance of named entity recognition in agricultural field, LSTM-CRF, BILSTM-CRF, CNN-BILSTM-CRF and BERT-BILSTM-CRF are selected to compare the same data set.

The comparison results are shown in Table 3.

Table 3: Performance comparison of each model

Model	P	R	F1
LSTM-CRF	81.51	81.33	81.44
BILSTM-CRF	84.37	83.47	83.92
CNN-BILSTM-CRF	85.11	84.86	84.98
BERT-BILSTM-CRF	87.46	86.03	86.74
PBERT-BILSTM-CRF	89.36	88.46	88.91

In order to verify the effectiveness of bilstm in improving data long-distance dependence, the comparative experiments of LSTM-CRF and BILSTM-CRF are carried out. From the results of accuracy, recall and F1 value, BILSTM model is improved by 2.86%, 2.14% and 2.48% respectively. Compared with LSTM, BILSTM can make full use of context information. In order to verify the influence of adding local features on the experiment, BILSTM-CRF and CNN-BILSTM-CRF are compared. The results show that the accuracy is improved by 0.74%, the recall is increased by 1.39%, and the F1 value is increased by 1.06%, indicating that CNN does improve the final result in extracting local features, And CNN has a very good recognition effect in including mixed Chinese and English characters and special characters. In order to verify the effectiveness of BERT model, the comparative experiments of CNN-BILSTM-CRF and BERT-BILSTM-CRF are carried out. The results show that these three indexes are improved. This shows that although CNN can fuse local features, BERT can find good vectors in flexible expression and word ambiguity to improve the ability of feature extraction. Finally, we compare the benchmark BERT and PBERT model. It can be found that the accuracy and recall of the pre trained BERT model are improved by 1.90% and 2.43% respectively, and reach the optimal F1 value of 88.91%.

In addition, the change of F1 value with iterative value during the experiment is drawn into a broken line diagram, as shown in Fig 5.

ISSN 2616-5767 Vol.5, Issue 4: 34-42, DOI: 10.25236/AJETS.2022.050407

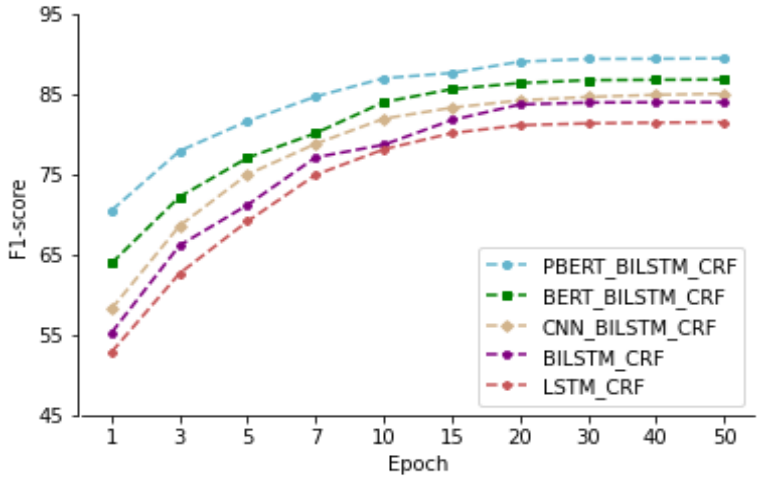

Figure 5: F1 value variation diagram of each model

It can be seen that PBERT model can reach a high level in the early stage of training, then continue to rise, and finally tend to be stable. It can be clearly seen that the F1 value of PBERT model is always better than other models. The initial value of other models is low, and it takes several rounds of iteration to reach a high level. After 50 rounds of iteration, each model tends to be stable, and the F1 value of the model proposed in this paper is the best.

4.3.2 Combined feature recognition

In recent years, the radical feature of Chinese characters has been widely used to enhance different Chinese natural language processing tasks. We also apply the radical feature to the model. In order to clearly and intuitively see the characteristics, we use Matplotlib to draw several representative partial radical thermodynamic diagrams of each entity category. As shown in Figure 6, 0, 1, 2 and 3 represent crop, disease, pest and drug entities respectively.

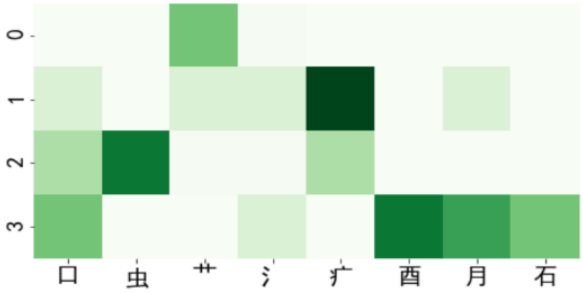

Figure 6: Thermodynamic diagram of representative partial distribution of different entities

For the four types of agricultural entities to be identified, the features of the corpus are selected according to different classifications according to the thermal color in Figure 6. Table 4 shows the text itself and the side of the text used in this paper as features. When recognizing the entity, the features are combined and combined with the context reference information to form different feature combinations.

Table 4: Custom feature collection

Features	Basis and Setting
Word	Crop varieties often have "瓜、菜、豆" and so on; Diseases often include "病,症" and so on; Pests often include "虫,蛾,螨" and so on; Pesticides often contain "素,胺,磷" and so on. These features are marked as 1 if any and 0 if none
Radical	Crops often have "艹、木、禾" and other side; Most of the diseases have "疒, 宀" and other laterals; Pests often include "虫、马,鸟" and so on; Pesticides often contain "口, 月,石" and so on. If it has such characteristics, it will be marked as 1-crop, 1-disc, 1-pest and 1-drug. If it is not marked as 0.

The experimental results of additional features are listed in Table 5.

Table 5: Experimental results after adding joint features

ISSN 2616-5767 Vol.5, Issue 4: 34-42, DOI: 10.25236/AJETS.2022.050407

Model	P	R	F1
PBERT-BILSTM-CRF	89.36	88.46	88.91
+ Word	88.94	88.47	88.70
+ Radical	89.97	89.54	89.75
Ensemble	90.32	90.16	90.24

As can be seen from the results in Table 5, the effect of adding a single word feature is not significant, but will reduce the F1 value. The reason may be that we define a single rule, and the common word feature leads to fuzzy recognition. After adding a single side feature, the accuracy, recall and F1 value are improved to a certain extent. Finally, after combining the word feature and the side feature, the joint feature model achieves the optimal F1 value of 90.24%.

4.3.3 Different entity recognition

In addition to observing the evaluation indicators of different models on the whole data set, we also carefully observed the performance of our model on different types of agricultural entities. The performance is shown in Table 6.

Table 6: Comparison of model recognition entity effects

Entity category	P	R	F1
Crop	93.68	94.73	94.20
Disease	91.73	90.98	91.35
Pest	91.36	91.24	91.30
Drug	82.11	83.76	82.93

The model performs well in the identification of crop names and pests, and the F1 scores are 94.20%, 91.35% and 91.30% respectively. It indicates that these entities are easy to identify correctly. However, the performance in pesticide identification is low, and the F1 value is 82.93%. This may be due to lack of boundary features and inconsistencies caused by the mixed use of characters, numbers and letters. Some categories of limited data may also affect performance.

5. Conclusions

Aiming at the NER task in the agricultural field, this paper establishes the data set in the agricultural field through network crawling, defines four main types of entities, and proposes a neural network structure model for pre training Bert. By comparing with other models and benchmark Bert model, the advantages of our model are proved. After adding joint features, our pbert-bilstm-crf model reaches the best F1 value of 90.24%. The fine tuning model based on pbert has the characteristics of fast training speed and high sharing degree, and can effectively identify common agricultural entities. At the same time, the performance of pesticide entity recognition is low. In the future work, we will improve the results by expanding the corpus and term dictionary in the field of agriculture. Entity recognition is a necessary work of knowledge atlas. In the future, we will focus on the task of relationship extraction. The construction of the model provides a basis for the construction of agricultural knowledge graph and the research of question and answer system in the future.

References

[1] ZHAO C J. Research on the development status and strategic objectives of smart agriculture [J]. China Agricultural Abstracts - agricultural engineering,2019.

[2] WANG C Y, WANG F. Research on agricultural named entity recognition based on conditional random field [J]. Journal of Hebei Agricultural University,2014,37(1):132-135. DOI:10.13320/j.cnki.jauh.2014.0027.

[3] LI J X, WANG P. A review of research methods of Chinese named entity recognition [J]. Computer Era,2021(4):18-21. DOI:10.16644/j.cnki.cn33-1094/tp.2021.04.005.

[4] Dong C,Zhang J,Zong Cet al. Character-Based LSTM-CRF with Radical-Level Features for Chinese Named Entity Recognition[J].Springer International Publishing,2016.

[5] WANG X M, TAO H C. Research on Chinese Named Entity Recognition Based on deep learning [J]. Journal of Chengdu University of information engineering,2020,35(3):264-270. DOI:10. 16836/j.cnki.jcuit.2020.03.003.

[6] Lample G,Ballesteros M,Subramanian S,et al.Neural Architectures for Named Entity

ISSN 2616-5767 Vol.5, Issue 4: 34-42, DOI: 10.25236/AJETS.2022.050407

Recognition[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.2016.

[7] ZHANG J, WU Q, YANG X Y,et al. Agricultural named entity recognition based on conditional random field [J]. Computer and modernization. 2018, No.269(01):123-126.

[8] LIU X J, GAO L C, SHI X Z. Named entity recognition based on Bi LSTM and attention mechanism[J]. Journal of Luoyang Institute of Technology (NATURAL SCIENCE EDITION) ,2019, 29(01):68-73+80.

[9] Qiang Z,Yong S B,Lz B,et al.Named entity recognition method in health preserving field based on BERT[J].Procedia Computer Science,2021,183:212-220.

[10] Cao, P. , Chen, Y. , Kang, L. , Zhao, J. , & Liu, S. . (2018). Adversarial Transfer Learning for Chinese Named Entity Recognition with Self-Attention Mechanism. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.

[11] ZHAO P F, ZHAO C J,WU H R, Named entity recognition of agricultural text based on attention mechanism[J]. Journal of agricultural machinery, 2021,52(1):185-192. DOI:10.6041/j.issn.1000-1298.2021.01.021.

[12] Devlin J,Chang M W,Lee K,et al.BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding[J].2018.

[13] Guo X,Zhou H,Su J,et al.Chinese agricultural diseases and pests named entity recognition with multi-scale local context features and self-attention mechanism[J].Computers and Electronics in Agriculture, 2020, 179(5):105830.

[14] Cw A,Hong W,Hui Z A,et al.Chinese medical named entity recognition based on multi-granularity semantic dictionary and multimodal tree[J].Journal of Biomedical Informatics,2020,111.

[15] X Li,Zhang H,Zhou X H.Chinese Clinical Named Entity Recognition with Variant Neural Structures Based on BERT Methods[J]. Journal of Biomedical Informatics, 2020, 107(5):103422.

[16] LI F L,KE J. Research progress of word vector semantic representation [J]. Information science,2019,037(005):155-165.

[17] Vaswani A,Shazeer N,Parmar N,et al.Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems.New York,2017:6000-6010.

[18] Hochreiter,Sepp,Schmidhuber,et al.Long short-term memory.[J].Neural Computation,1997.

[19] Luo, L. , Yang, Z. , Yang, P. , Zhang, Y. , Wang, L. , & Lin, H. , et al. (2017). An attention-based bilstm-crf approach to document-level chemical named entity recognition. Bioinformatics(8), 8.

[20] Xu G H,WANG c Y,HE X F.Improving clinical named entity recognition with global neural attention[c]//Asia PacificWeb and Web·Age Informa“on Management Joint International Conference on Web and Big Data,2018:264—279.

[21] LI Ni,GuAN Huanmei,YANG Piao,et a1.BERT—IDcNN—CRF for named entity recogni“on in chinese[J]. Journal ofshandong university(Natural science),2020,55(1):102-109.