agriculture

MDPI

Article

AWdpCNER: Automated Wdp Chinese Named Entity Recognition from Wheat Diseases and Pests Text

Demeng Zhang [1], Guang Zheng [1,2], Hebing Liu [1], Xinming Ma [1,2] and Lei Xi [1,2,*]

[1] College of Information and Management Sciences, Henan Agriculture University, Zhengzhou 450046, China; zdmeng@stu.henau.edu.cn (D.Z.)
[2] Henan Engineering Laboratory of Farm and Monitoring and Control, Zhengzhou 450002, China
* Correspondence: xil@henau.edu.cn; Tel.: +86-138-0386-6921

Abstract: Chinese named entity recognition of wheat diseases and pests is an initial and key step in constructing knowledge graphs. In the field of wheat diseases and pests, there are problems, such as lack of training data, nested entities, fuzzy entity boundaries, diverse entity categories, and uneven entity distribution. To solve the above problems, two data augmentation methods were proposed to expand sentence semantic information on the premise of fully mining hidden knowledge. Then, a wheat diseases and pests dataset (WdpDs) for Chinese named entity recognition was constructed containing 21 types of entities and its domain dictionary (WdpDict), using a combination of manual and dictionary-based approaches, to better support the entity recognition task. Furthermore, an automated Wdp Chinese named entity recognition model (AWdpCNER) was proposed. This model was based on ALBERT-BiLSTM-CRF for entity recognition, and defined specific rules to calibrate entity boundaries after recognition. The model fusing ALBERT-BiLSTM-CRF and rules amendment achieved the best recognition results, with a precision of 94.76%, a recall of 95.64%, and an F1-score of 95.29%. Compared with the recognition results without rules amendment, the precision, recall, and F1-score was increased by 0.88 percentage points, 0.44 percentage points, and 0.75 percentage points, respectively. The experimental results showed that the proposed model could effectively identify Chinese named entities in the field of wheat diseases and pests, and this model achieved state-of-the-art recognition performance, outperforming several existing models, which provides a reference for other fields of named entities recognition such as food safety and biology.

Keywords: Chinese named entity recognition; wheat diseases and pests; data augmentation; ALBERT-BiLSTM-CRF; rules amendment

Citation: Zhang, D.; Zheng, G.; Liu, H.; Ma, X.; Xi, L. AWdpCNER: Automated Wdp Chinese Named Entity Recognition from Wheat Diseases and Pests Text. *Agriculture* **2023**, *13*, 1220. https://doi.org/10.3390/agriculture13061220

Academic Editor: Luís Manuel Navas Gracia

Received: 28 April 2023
Revised: 2 June 2023
Accepted: 8 June 2023
Published: 9 June 2023

1. Introduction

Wheat is one of the world's most important food crops. In the process of wheat planting, there is a variety of diseases and pests, and their occurrence directly affects the yield and quality of wheat. Currently, a wealth of knowledge on wheat disease and pest prevention and control measures is often stored in the form of unstructured text in web pages, books, and literature. The storage, organization, representation, and management of this data varies greatly, which leads to dispersed and confusing data in this field, making it difficult for people to quickly access accurate information on disease and pest control and leaving them unable to carry out precise prevention. The construction of a knowledge graph in the field of wheat diseases and pests and the representation of disease, pest, and control data in a structured form could help people locate valuable information efficiently and accurately, which is of great significance for precise control. The named entity recognition of wheat diseases and pests is a key step in the construction of a knowledge graph in the field. It aims to identify and classify relevant named entities from these unstructured data. The quality of a knowledge graph is directly determined by the recognition results [1,2].

In recent years, named entity recognition had been widely used in vertical fields. With the development of deep learning, methods based on deep learning had become the

mainstream model for Chinese named entity recognition. Liu XL et al. (2021) [3] proposed the BERT-CRF model to realize named entity recognition in the field of the raw egg supply chain, with a precision of 91.82%. Yang YL et al. (2021) [4] combined BiLSTM and CRF to realize named entity recognition in the field of TCM cases. Xu L et al. (2021) [5] proposed named entity recognition in the biomedical field based on the BERT-BiLSTM-CRF model, which effectively solved the problem of low accuracy in the semantic recognition of static word vector representation. Shen TP et al. (2022) [6] proposed the BERT-BiLSTM-CRF model and achieved excellent results in both MSRA and People's Daily Corpus. All the above models achieved good recognition results when the training corpus of their fields was sufficient, but they could not be directly applied to the field of wheat diseases and pests, with an insufficient corpus.

In the field of agriculture, Malarkodi et al. (2016) [7] and Guo Xet al. (2020) [8] extracted local features of texts through CNN, and then combined this with the BiLSTM+CRF model to realize named entity recognition in the field of agricultural pests and diseases. Yan LH et al. (2021) [9] and Yu HL et al. (2021) [10] used the BiLSTM+CRF model to realize the named entity recognition of grape and rice diseases and pests, respectively. The methods in the above literature [7–10] achieved good results, but the traditional word vector model is easily affected by the results of word segmentation and cannot represent the polysemy phenomenon. Li Y (2021) [11], Ren N et al. (2021) [12], and Zheng YZ et al. (2021) [13] combined the BERT pre-trained language model with the BiLSTM+CRF model to realize named entity recognition in the field of agricultural pests and diseases. In the literature [11–13], BERT was used to replace the traditional word vector model, which effectively reduced the impact of word segmentation errors and solved the polysemy problem of a word. However, the BERT model depends on millions of parameters, which are time-consuming and costly to train.

The vertical domain methods proposed in references [3–13] provided a guideline for carrying out entity recognition in the field of wheat diseases and pests in this paper. However, these studies usually only identified entity categories such as diseases, pests, varieties, drugs, and harmful organs, and the recognition effect was better when there were few categories. At present, in the study of named entity recognition in the field of agricultural plant diseases and pests, the research on the real corpus is lacking. There are some problems in this field, such as the lack of training data, nested entities, fuzzy entity boundaries, diverse entity categories, and uneven entity distribution. The above entity categories could not fully extract the information implied in the unstructured text, which was insufficient to explain the problem of named entity recognition in the field of agricultural diseases and pests. To solve the problem of named entity recognition in the field of wheat diseases and pests, this paper constructed the dataset WdpDs, and the dictionary WdpDict, for wheat diseases and pests. A model AWdpCNER combining deep learning with fusion rules is also proposed in this paper. This model adopts a strategy of combining the lightweight dynamic word vector model ALBERT with the BiLSTM-CRF model. Aiming at the problem of fewer entity categories, such as certain pathogens and wheat areas, we proposed two data augmentation methods to extend the sentence semantic information by similar word substitution to make up for the lack of training data, which effectively improved the results of named entity recognition of wheat diseases and pests under the condition of small samples. Specific rules were defined for special entities with fuzzy boundaries, such as drugs and symptoms, to align entity boundaries, to further improve the overall recognition results of the model, and to provide support for downstream tasks such as the construction of a knowledge graph and the knowledge question answering of wheat diseases and pests.

2. Materials and Methods

2.1. Dataset Construction and Characteristic Analysis

Aiming at the research of Chinese named entity recognition of wheat diseases and pests, this paper constructed the entity recognition dataset WdpDs through two steps: corpus collection and pre-processing, and corpus annotation.

2.1.1. Corpus Collection and Pre-Processing

To ensure the correctness and reliability of the training data, the data of wheat diseases and pests were mainly derived from two places: the first of these was two authoritative books—*Identification and Control of Wheat diseases and pests* [14] and *Atlas of Diagnosis and Control of Wheat diseases and pests* [15]. Secondly, crawlers were used to grab data from authoritative websites such as *China Crop Germplasm Information System*, *National Agricultural Science Data Center* and *Baidu Wikipedia*. Firstly, the OCR recognition algorithm was used to convert the two books into electronic and text format, and to manually modify the wrong words and garbled characters. Secondly, spaces, blank lines, and special symbols were manually removed from web page data. Finally, duplicate data and invalid data were removed. Ultimately, a dataset containing more than 7000 samples of diseases, pests, and control measures was constructed, with about 250,000 characters.

2.1.2. Corpus Annotation

Under the guidance of field experts, combined with a summary of existing research on entity recognition in the agricultural field, the characteristics of the data of wheat diseases and pests were deeply analyzed, the implicit knowledge was fully mined, and the entity categories of wheat diseases and pests were carefully divided into 21 categories—including disease, disease class, pest, pest class, pest time cycle, pathogeny, pathogeny class, wheat organ, drug, agricultural control, wheat growth time, wheat variety, wheat area, symptom, organ symptom, harmful crop, harmful area, genus, family, other name, and enemy—to ensure the integrity of the entity categories of WdpDs.

The raw corpus was labeled using the BIO scheme. To reduce the labeling time and ensure the consistency of entity labeling, corpus labeling adopted the automatic labeling method of combining dictionary and manual. Firstly, common entities were extracted during data pre-processing to construct the domain dictionary WdpDict. Secondly, according to the domain dictionary WdpDict, the characters of the raw corpus were matched to realize automatic entity labeling. Finally, the corpus was manually adjusted and improved, and the WdpDict was dynamically updated in the process. The specific tagging process is shown in Figure 1.

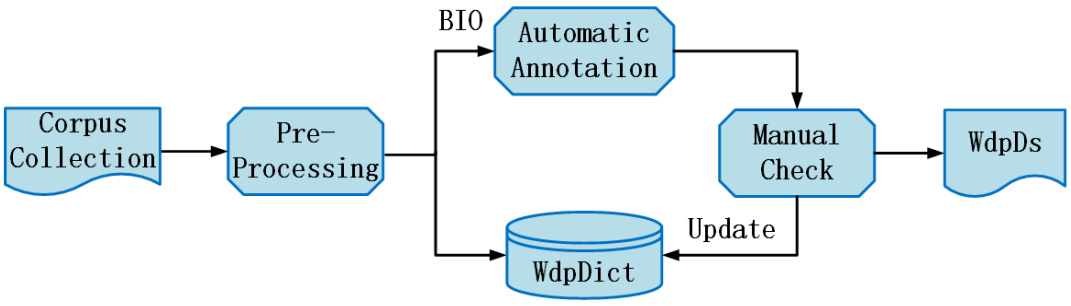


Figure 1. Tagging Process of Wheat Diseases and Pests Data Set.

After the above processing, the wheat disease and pests entity recognition dataset, WdpDs, was finally constructed, which contained 18,127 entities of 21 entity categories. The number distribution of each entity category is shown in Table 1. Examples of annotations are shown in Figure 2. Where B- represents the starting position of the entity, I- represents the middle or end position of the entity, and O represents the non-entity.

Table 1. Entity category details of wheat diseases and pests dataset.

ID	Entity Name	Entity Tags	Number of Entities	Examples
1	Disease	DIS	2539	powdery mildew, scab
2	Disease class	DIS_CLA	173	Fungal disease, nematode disease
3	Pest	PES	2090	aphids, armyworm
4	Pest class	PES_CLA	156	Underground pests, leaf pests
5	Pests time cycle	PES_TIM	1619	adults, nymphs
6	Pathogeny	PAT	407	brucella gramineae
7	Pathogeny class	PAT_CLA	311	fungi, viruses
8	Wheat organ	ORG	2363	leave, stem
9	Drugs (Termiticides)	DRU	1218	triadimefon powder
10	Agricultural control	CON	439	resistant varieties, watering
11	Wheat growth time	TIM	601	jointing stage, grouting stage
12	Wheat variety	WHE	456	yumai 18, jimai38
13	Wheat area	ARE	276	northwest spring wheat area
14	Symptom	SYM	1036	dry death
15	Organ symptoms	OSYM	1837	yellowing of the leave
16	Harmful crops	CROP	1066	Wheat, corn
17	Harmful region	REG	701	Henan, Zhengzhou
18	Genus	GEN	203	hemiptera, lepidoptera
19	Family	FAM	201	noctuidae, culicidae
20	Other name	OTHN	240	Wheat ear dry, oil worm
21	Enemy	ENE	195	Seven-star ladybird, hoverfly

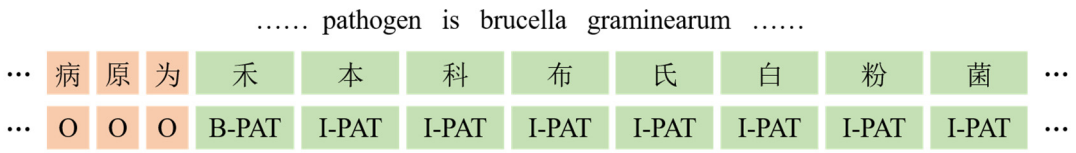


Figure 2. Tagging example of BIO.

2.1.3. Analysis of Corpus Characteristics

Through a comprehensive analysis of the wheat diseases and pests dataset, WdpDs, we found that the domain entities were mainly characterized by the following four aspects:

1. The boundary features of some entities in the WdpDs dataset were not obvious, which meant they were easily broken down incorrectly. For example, "3% clozophos (Mille) granules" was a typical example.
2. The entity structure of wheat diseases and pests was complex, and some entities were composed of numbers, letters, and Chinese characters. For example, "Xinong 6082", "75% Malathion oil", and some other entities were typical examples.
3. There was nesting among some entities in the corpus of wheat diseases and insect pests. For example, the pathogenic entity "wheat yellow dwarf virus" was nested in the disease entity "wheat yellow dwarf disease", and so on.
4. The dataset of wheat diseases and pests contained many entity categories. The constructed WdpDs dataset contained 21 types of entities, more than the JE-DPW dataset in the same domain [16].

2.2. AWdpCNER Model

In this paper, the AWdpCNER model was used to identify the named entities of wheat diseases and pests. The model adopted the strategy of ALBERT-BiLSTM-CRF combined with rules amendment. The overall architecture of the model is shown in Figure 3. The steps of the AWdpCNER model can be found in Appendix A.

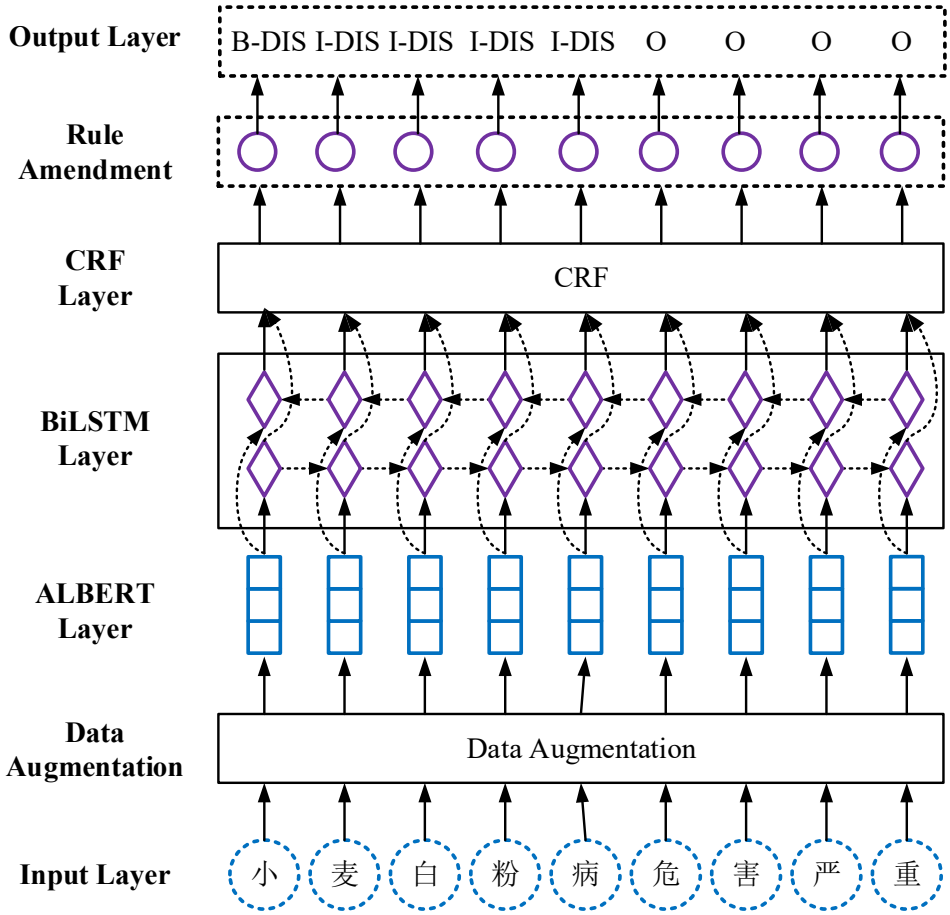


Figure 3. Overall identification framework of wheat diseases and pests.

The whole network can be divided into seven layers: input layer, data augmentation, ALBERT layer, BiLSTM layer, CRF layer, rule amendment, and output layer. Firstly, the semantic information of the entity categories with a small number of entities, such as Pathogen and Wheat Area, was expanded by data augmentation, and then the dataset, constructed through a combination of dictionary and manual semi-automatic annotation, was used as an input to the ALBERT layer. Secondly, the lightweight pre-training model ALBERT was used to generate dynamic word vectors containing context information, which effectively alleviated the polysemy problem. Meanwhile, to improve the accuracy of the output features of the ALBERT layer, the word vector was input to the BiLSTM layer to further model the context features. Finally, the sequence labels' outputs by the BiLSTM layer were constrained and modified by CRF and its rules, and the final predicted label sequence was obtained.

2.2.1. Data Augmentation

In recent years, entity recognition methods based on deep learning have been widely used in many fields. However, deep learning models often require a large amount of training data. In the field of wheat diseases and pests, due to the lack of training data, the complex entity structure, the diverse entity types, and the uneven distribution of entities, the research of named entity recognition in this field has certain challenges. To address the above problems, this paper proposed two methods of data augmentation (DA), the main idea of which was to supplement the semantic information of sentences and to compensate for the lack of training data.

1. Data augmentation Method 1 (DA1): Under the condition of maximum guarantee of sentence sequence integrity, the text paragraphs in the original dataset were randomly shuffled, and the shuffled paragraphs were copied back to the original dataset.

2. Data augmentation Method 2 (DA2): An entity was randomly selected from the wheat diseases and pests text data, and then a synonym of the entity was randomly selected from the constructed domain dictionary, WdpDict, for replacement, and the replaced text data was copied back to the original dataset.

This study mainly focused on augmenting the entity categories with fewer instances, such as disease class, pest class, and pathogeny class. In the end, we obtained a total of 21,345 entities through data augmentation.

2.2.2. ALBERT Layer

In the field of NLP, the commonly used language models for transforming text data into word vectors include Word2Vec [17], GloVe (Global Vectors for Word Representation) [18], One-Hot, etc. However, the word vectors trained by the above models are static and cannot represent the polysemy phenomenon.

BERT is a pre-trained language model proposed by Google. It is a bidirectional encoder network constructed based on the Transformer [19] neural network, which cannot only obtain word-level features containing context information, but also effectively capture sentence-level features [20]. Compared with the traditional word vector model, the word vector trained by the BERT model was based on the contextual information to generate a dynamic word vector, which effectively solved the problem of polysemy. However, despite the excellent performance of BERT in various tasks, the number of parameters reached 108M, which required a large-scale corpus for training and costs a lot.

To solve the problem of the number of BERT parameters, Lan et al. (2019) [21] proposed a lightweight pre-trained language model, ALBERT (A Lite BERT), which was like BERT in model structure, but the number of parameters was only one ninth of those of BERT. Based on ensuring the performance of the BERT model, the model also made the following three improvements, which greatly reduced the space occupied by the model and significantly improved the training speed.

1. Factor word embeddings: In the ALBERT model, the one-hot vector was mapped to a low-dimensional space first, and then to the hidden layer. The complexity transformation of the parameter number calculation from the BERT model to the ALBERT model is shown in Equation (1):

$$\mathrm{O}(V \times H) \rightarrow \mathrm{O}(V \times E + E \times H) \tag{1}$$

where V is the vocabulary length, H is the hidden layer dimension, and E is the word embedding dimension. $E = H$ in BERT, and $E \ll H$ in ALBERT.
2. Cross-layer parameter sharing: In ALBERT, the parameters were shared in both the full connection layer and the attention layer, that is, all parameters in the Encoder were shared, which greatly reduced the number of model parameters and improved the training speed, but the reduced number of model parameters also degraded its performance.
3. Sentence Order Prediction: In order to compensate for the performance loss caused by the reduction in the number of parameters, ALBERT proposed the inter-sentence coherence prediction SOP (sentence order prediction) to improve the model performance. Different from the original NSP (next sentence prediction) task of the BERT model, SOP removed the influence of topic prediction and only preserved the relational consistency prediction.

2.2.3. BiLSTM Layer

The encoder of ALBERT partially adopted a self-attention mechanism, which led to inaccurate relative position information being extracted and a lack of sequence in the output features. Therefore, this paper adopted the BiLSTM model to further model the context features.

LSTM (long short-term memory) [22] was improved on the basis of RNN, and effectively solved the problem of gradient explosion or gradient disappearance in long sequence text. However, one-way LSTM could only learn forward information, but not backward information. Therefore, Graves et al. (2005) [23] proposed BiLSTM (bidirectional long short-term memory) composed of forward LSTM and backward LSTM. The basic idea of BiLSTM was to carry out forward propagation and backward propagation respectively for each word in the sequence, and then connect the results to the output. Thus, it could better capture the bidirectional information of long sequence text. The structure of the BiLSTM model is shown in Figure 4.

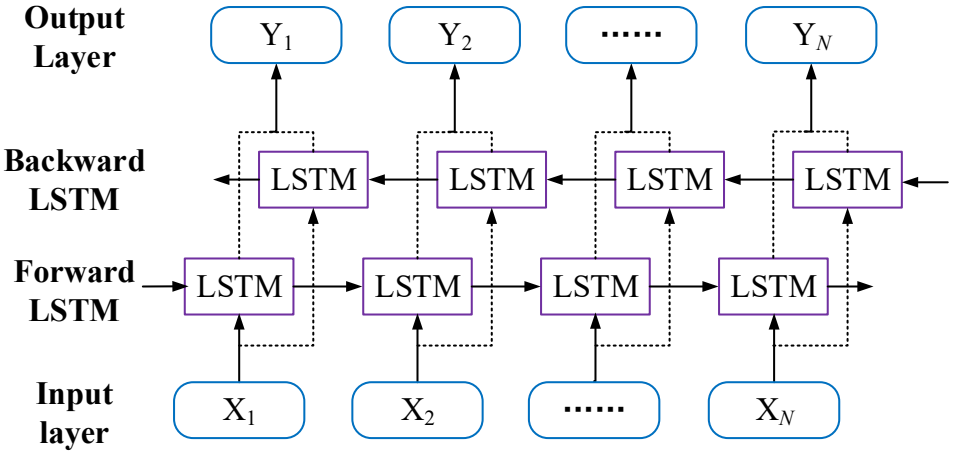


Figure 4. Bidirectional long short-term memory model diagram.

2.2.4. CRF Layer

Although the BiLSTM layer could further learn the context features, it did not consider the dependencies between neighboring labels and always chose the label with the highest probability as the output, which might lead to B-label1 followed by I-label2. Since the CRF model could learn the dependencies between neighboring labels, CRF was introduced after BiLSTM layer to improve the accuracy of model prediction in this paper.

CRF was first proposed by Lafferty et al. (2002) [24] and was mainly used for sequence annotation. In the process of model training, CRF could automatically learn the constraints between sentences and obtain the label transition probability, so as to ensure the legitimacy of the predicted labels and reduce the wrong prediction sequences. The specific constraints of CRF were as follows:

1. The first word in a sentence always begins with the label "B-" or "0", not "I-".
2. In the label "B-label1 I-label2 I-label3 I-...", label1, label2, and label3, should belong to the same entity. For example, "B-DIS I-DIS" was a valid label sequence, while "B-DIS I-DRU" was an invalid label sequence.
3. The first label of the entity should start with "B-, not "I-". For example, "O I-DIS" was a valid label sequence, while "O I-DIS" was an invalid label sequence.

In this study, the CRF layer mainly performed conditional constraints on the label sequences output by the BiLSTM layer, so as to obtain reasonable sequences with maximum probability.

2.2.5. Rules Amendment

By analyzing the characteristics of the data in this field and the errors in the recognition process, the following three types of rules were formulated to further improve the recognition effect of the model:

1. For pest and disease entities, if harmful crops appear before them, they shall be labeled as a whole. For symptom entities, if organs appear in the adjacent vocabulary, the whole entity is modified to organ symptom type entity. In the process of rule

correction, a sliding window with the size of 1 was set, centering on the keyword to search an entity for the context. If the adjacent prediction labels were entity terms, the corresponding rules were found and merged into related entities. Otherwise, the original word prevailed. The specific rules are shown in Table 2.

Table 2. Pest and symptom entity rule definitions.

Rule Definition	Entity Examples
CROP + DIS = DIS	Wheat powdery mildew, gibberellic disease of corn
CROP + PES = PES	Wheat tube thrips, corn aphids
ORG + SYM = OSYM	Stem dry, leaves yellow

2. Diseases often ended with the word "disease". The last word of this type of entity was concatenated with the next word immediately adjacent to it. If a whole word could be formed, it would be regarded as the whole prediction. The drugs were usually composed of their concentration and drug name. Regex was written to recognize numbers, symbols, and Chinese as a whole. The specific rules are shown in Table 3.

Table 3. Disease and drug entity rule definitions.

Entity Label	Sentence	Rules	Results
DIS	Wheat blue dwarf disease virus belongs to virus Stripe rust disease attacks wheat	disease [\u4e00-\u9fa5]	disease virus/n disease attacks/empty
DRU	50% phoxim emulsion oil 20% kiku·horse emulsion	\d+(?:\.\d+)?(?:%)(?:[\u4e00-\u9fa5]+·)?(?=[\u4e00-\u9fa5])	50% 20% kiku·

3. All the predictions of the AWdpCNER model are amended, such as the wrong prediction that label1 and label2 are different types of entities in "B-label1 I-label2" and the beginning of "I-label".

3. Results

3.1. Experimental Parameters Setup

The hardware environment that the experimental research relied on was Intel(R) Xeon(R) Silver4116 CPU@2.10GHz. The software environment was Python3.6 and tensorflow1.14. The model parameters were set as follows: based on ALBERT_Base version, with 12 transformer layers, 768 hidden layers, and 12 multi-head attention mechanisms. The maximum sequence length was 256, the BiLSTM contained 256 dimensional hidden layers, the learning rate was 5×10^{-5}, the training batch_size was 64, the dropout was 0.5, the clip was 0.5, the optimizer chose Adam, and the number of iterations was 100. The model recognition results were evaluated by Precision (P), Recall (R), and F1-score($F1$). The specific formulas are shown in Formulas (2)–(4):

$$P = \frac{TP}{TP + FP} \times 100\%, \tag{2}$$

$$R = \frac{TP}{TP + FN} \times 100\%, \tag{3}$$

$$F1 = \frac{2 \times P \times R}{P + R} \times 100\%, \tag{4}$$

where TP represents the number of positive samples with correct prediction, FP represents the number of positive samples with incorrect prediction, and FN represents the number of negative samples with incorrect prediction.

3.2. Experiment Results

The constructed dataset, WdpDs, was divided into training set, testing set, and validation set according to the ratio of 8:1:1. The identification results of different models were compared according to the three evaluation metrics proposed in Section 3.1.

3.2.1. Performance Comparison of Different Models

For the divided training set, testing set, and validation set, four groups of models, Word2Vec-IDCNN-CRF (2017) [25], Word2Vec-BiLSTM-CRF [26], BERT-BiLSTM-CRF [27], and ALBERT-BiLSTM-CRF, were set to conduct experiments, respectively. The test results are shown in Table 4.

Table 4. Results for each model.

Model	P/%	R/%	F1-Score/%
Word2Vec-IDCNN-CRF	85.49	87.29	86.38
Word2Vec-BiLSTM-CRF	88.05	89.4	88.72
BERT-BiLSTM-CRF	90.9	91.16	91.03
ALBERT-BiLSTM-CRF	90.86	91.70	91.28

As can be seen from Table 4, the recognition effect based on Word2Vec-BiLSTM-CRF was significantly better than that of Word2Vec-IDCNN-CRF, because IDCNN could only obtain local features, while BiLSTM could obtain global features. In the case of long text sequences, BiLSTM had a better recognition effect. The P, R, and F1 of the model were increased by 2.56 percentage points, 2.11 percentage points, and 2.34 percentage points, respectively. Using BiLSTM-CRF as the baseline model, the two types of vectors embedding models Word2Vec and BERT were compared. It can be found from Table 4 that P, R, and F1 based on the BERT model were increased by 2.85 percentage points, 1.76 percentage points, and 2.31 percentage points, which proved that BERT could effectively represent polysemy and improve the model recognition effect.

The comparison between the BERT and ALBERT pretrained language models showed that the overall performance of the ALBERT model was better, with the R and F1 increased by 0.54 percentage points and 0.25 percentage points, respectively. In addition, the training time of the BERT-BiLSTM-CRF model for 100 epochs was 26.45 h, while the training time of ALBERT-BilSTM-CRF was 19.01 h, which indicated that the reduction in the number of parameters could significantly improve the training speed of the ALBERT model.

3.2.2. Comparison of Results of Different Data Augmentation Methods

According to the test results of the four groups of models in Section 3.2.1, the ALBERT-BiLSTM-CRF model achieved the highest F1-score of 91.28% in the wheat diseases and pests dataset, WdpDs. Based on this model, two data augmentation methods to expand the scale of dataset, WdpDs, are introduced in this section. The experimental results after augmentation are shown in Table 5.

Table 5. Results for model after data augmentation.

Model	P/%	R/%	F1-Score/%
ALBERT-BiLSTM-CRF	90.86	91.70	91.28
DA1 + ALBERT-BiLSTM-CRF	92.85	93.14	92.99
DA2 + ALBERT-BiLSTM-CRF	91.46	92.31	91.88
DA1 + DA2 + ALBERT-BiLSTM-CRF	93.88	95.2	94.54

As can be seen from Table 5, both data augmentation methods can improve the overall recognition results of the model, and the F1-score was increased by 1.71 percentage points and 0.6 percentage points, respectively, indicating that DA1 had a greater improvement on the model performance. After combining the two data augmentation methods, the

ALBERT-BiLSTM-CRF model achieved the highest P, R, and F1 on the extended WdpDs, which were 93.88%, 95.2%, and 94.54%, respectively. Compared with the ALBERT-BiLSTM-CRF model without data augmentation, its P, R, and F1-score improved by 3.02 percentage points, 3.5 percentage points, and 3.26 percentage points, respectively. The results indicate that enriching the sample through data augmentation can, to a certain extent, optimize the overall recognition performance of the model.

3.2.3. Entity Recognition Results for AWdpCNER Model

After augmenting the original dataset, WdpDs, by combining the two data augmentation methods, the identification of named entities of the WdpDs dataset were carried out based on the ALBERT-BiLSTM-CRF model. The specific precision, recall, and F1-score of the 21 types of entities are shown in Table 6.

Table 6. Results for all entities.

ID	Entity	P/%	R/%	F1-Score/%
1	Disease (DIS)	95.24	93.33	94.28
2	Disease class (DIS_CLA)	88.89	100	94.12
3	Pest (PES)	92.77	95.06	93.9
4	Pest class (PES_CLA)	85.71	100	92.31
5	Pest time cycle (PES_TIM)	98.49	97.51	98
6	Pathogeny (PAT)	76.79	87.76	81.9
7	Pathogeny class (PAT_CLA)	87.5	91.3	89.36
8	Wheat organ (ORG)	91.32	93.08	92.19
9	Drug (DRU)	89.6	89.6	89.6
10	Agricultural control (CON)	79.01	82.05	80.5
11	Wheat growth time (TIM)	92.86	98.48	95.59
12	Wheat variety (WHE)	91.3	97.67	94.38
13	Wheat area (ARE)	87.5	100	93.33
14	Symptom (SYM)	85.31	93.13	89.05
15	Organ symptom (OSYM)	84.09	82.96	83.52
16	Harmful crop (CROP)	93.2	91.4	92.29
17	Harmful region (REG)	93.18	93.89	93.54
18	Genus (GEN)	100	100	100
19	Family (FAM)	100	100	100
20	Other name (OTHN)	89.47	85	87.18
21	Enemy (ENE)	76.47	89.66	82.54

In general, the experimental results showed that the overall recognition performance of the model was good. As can be seen from Table 6, the F1-score of eight types of entity, including Pathogeny, Pathogeny Class, Drug, Agricultural Control, Symptom, Organ Symptom, Other Name, and Enemy, were all below 90%, and the precision of Pathogeny and Enemy are below 80%. Due to the small number of Pathogeny, Other Name, and Enemy, which account for a relatively low proportion of the dataset, the model could not fully learn their contextual features. The five types of Pathogen Class, Drug, Agricultural Control, Symptom, and Organ Symptom have longer lengths, flexible and diverse entity compositions, and complex structures, making model recognition difficult. The recall of Disease Class, Pest Class, Wheat Area, Genus, and Family was 100%, because these five entities all have clear boundary characteristics, ending with the words "pest", "worm", "area", "genus", and "family", respectively, and the precision of Genus and Family was also 100%. The evaluation indicators of P, R, and F1 for the other eight entities are all above 91%.

3.2.4. Comparison of Entity Recognition Results before and after Rules Amendment

According to the recognition results in Section 3.2.2, the DA1+DA2+ ALBERT-BiLSTM-CRF model had the best identification result with precision, recall, and F1-score of 93.88%, 95.2%, and 94.54%. On this basis, the rules formulated in Section 2.2.5 were added to further

correct the prediction results of four entities: disease, pest, drug, and organ symptom, in order to optimize the overall recognition performance of the model.

After rules amendment, the precision reached 94.76%, the recall reached 95.64%, and the F1-score reached 95.29%. Compared with the results of DA1 + DA2 + ALBERT-BiLSTM-CRF, the precision, recall, and F1-score of the model increased by 0.88 percentage points, 0.44 percentage points, and 0.75 percentage points, respectively, indicating that rules amendment had, to some extent, improved the model's performance. The specific recognition results of four types of entities are shown in Table 7. Table 8 shows some examples before and after rules amendment.

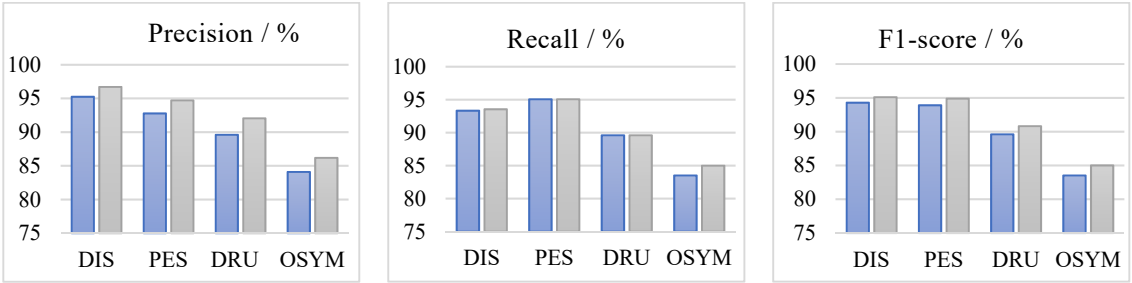
Table 7. Recognition results of DA1 + DA2 + ALBERT-BiLSTM-CRF model with fusion rules.

ID	Entity Name	P/%	R/%	F1-Score/%
1	Disease (DIS)	96.71	93.54	95.1
3	Pest (PES)	94.71	95.06	94.88
9	Drugs (DRU)	92.05	89.6	90.81
15	Organ symptoms (OSYM)	86.18	83.86	85

Table 8. Example results before and after rule amendment.

Sentence	Before Rule Amendment	After Rule Amendment
Wheat powdery mildew is a type of . . .	powdery mildew	Wheat powdery mildew
. . . leaves turned yellow . . .	leaves	leaves turned yellow
. . . 20% Chrysanthemum · Horse Emulsion . . .	Horse Emulsion	20% Chrysanthemum · Horse Emulsion

The comparison of the results before and after the rule amendment of Disease, Pest, Drug, and Organ Symptom are shown in Figure 5. As can be seen from Figure 5, after amending the definition rules of the ALBERT-BiLSTM-CRF model, the recognition effect of the four types of entities was improved to varying degrees, with precision having increased by 1.47 percentage points, 1.94 percentage points, 2.45 percentage points, and 2.09 percentage points, and with F1-score having increased by 0.82 percentage points, 0.98 percentage points, 1.21 percentage points, and 1.48 percentage points, respectively; organ symptoms had the most obvious improvement.



(**a**) precision comparison (**b**) recall comparison (**c**) F1-score comparison

Figure 5. Comparison chart of identification results before and after rule amendment. Blue represented the result before amendment, and gray represented the result after amendment.

4. Discussion

In recent years, with the development of deep learning technology, Chinese named entity recognition methods, based on deep learning, have gradually become mainstream and widely applied in various vertical fields. Entity recognition in the field of wheat diseases and pests is a key step in building the knowledge graph of the domain knowledge, and the quality of the knowledge graph is directly determined by the recognition results.

In the field of wheat diseases and insect pests, due to its complex text structure, diversified expressions, and its large number of domain proper noun and special symbols, entity

recognition research in this field is more challenging. Not all existing entity recognition methods are applicable to this field. Although researchers have conducted corresponding research on entities in this field, there are still problems, such as limited entity categories and insufficient knowledge mining. This article used two data augmentation methods to expand the number of entities and alleviate the impact of uneven entity distribution on recognition results. At the same time, a semi-automatic dataset, WdpDs, was constructed through a combination of the domain dictionary and manual work, which includes 21 entity categories and a total of 18,127 entities, with more refined entity segmentation. Compared with general datasets, the classification of entity categories is more refined. According to the text characteristics of wheat diseases and pests, this paper proposed the AWdpCNER model. The experimental results showed that, compared with other NER models, this model could better identify named entities in the field of wheat diseases and pests, and provided a reference for the recognition of named entities in other fields, such as food safety and biology.

Although this study performed excellently in the task of named entity recognition in the field of wheat diseases and pests, it can only recognize predefined entity categories. In future research, it is necessary to further explore methods for automatically discovering hidden entity categories from unstructured text in open domains.

5. Conclusions

1. Aimed at the problems of Chinese named entity recognition in the field of wheat diseases and pests, including the lack of training data, many proper nouns, diverse entity categories, and uneven entity distribution, the AWdpCNER model was proposed. The model combined two data augmentation methods to expand the semantic information of sentences, which improved the accuracy of the model for a small number of entity categories, and effectively solved the problem of named entity recognition in the case of small samples. The model recognition precision is 94.76%, the recall is 95.64%, and the F1-score is 95.29%.
2. The dynamic word embedding vector was obtained based on the lightweight ALBERT model, which could capture the entity context features, enrich the semantic representation of wheat disease and pest text, effectively alleviate the problem of polysemy representation, and improve the model recognition performance.
3. The specific rules were defined to modify the prediction results of the ALBERT-BiLSTM-CRF model. The experiment proved that the rule amendment alleviated the problems of fuzzy entity boundary and nesting among entities and, to a certain extent, optimized the model performance.

Author Contributions: Conceptualization, G.Z., H.L. and L.X.; data curation, G.Z.; formal analysis, D.Z.; funding acquisition, X.M.; investigation, D.Z.; methodology, D.Z.; project administration, L.X.; resources, D.Z.; supervision, G.Z.; validation, D.Z.; writing—original draft, D.Z.; writing—review and editing, L.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Modern Agricultural Industrial Technology System of Henan Province (S201001G04) and the National Key Research and Development Program (2016YFD0300609).

Data Availability Statement: No data about this paper are available due to privacy or ethical restrictions.

Acknowledgments: We thank the editors of Agronomy and the anonymous reviewers for their valuable suggestions.

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

Appendix A

As shown in Algorithm A1, the steps of our proposed AWdpCNER model can be summarized as follows:

Algorithm A1 The pseudocode of wheat disease and pest named entity recognition task

Input: wheat disease and pest sentence S, and its ground-truth labels Y.
Output: the predicted entity labels, the best training weight of the model

```
Loading all sentence in the field of wheat disease and pest
Data augmentation of sentences, and dividing training set, testing set, and validation set
Building the NER model, converting the above three sets into the dynamic embeddings
    E1, E2, E3 by ALBERT; further learning context features from E1 using BiLSTM, and using
    cross entropy loss function to optimize; calculating the transfer probability of labels using CRF,
    and obtaining the overall F1-score for predicted labels
For i in 100 epoch:
    For batch in train_iter:
        Using training set to run the NER model, and obtaining the weight of the model
    Evaluating the NER model using test and validation set, and obtain F1-score
    If F1-max > F1-score then F1-max ← F1-score, and save the weight of the model
Amending the above predicted labels according to the rules defined in Section 2.2.5, and
    recalculating F1-score
End for
```

References

1. Ren, Y.; Yu, H.; Yang, H.; Liu, J.; Yang, H.; Sun, Z.; Zhang, S.; Liu, M.; Sun, H. Recognition of quantitative indicator of fishery standard using attention mechanism and the BERT+BiLSTM+CRF model. *Trans. Chin. Soc. Agric. Eng. (Trans. CSAE)* **2021**, *37*, 135–141.
2. Wang, Y.; Zhang, C.; Bai, F.; Wang, Z.; Ji, C. Review of Chinese Named Entity Recognition Research. *J. Front. Comput. Sci. Technol.* **2023**, *17*, 324–341.
3. Liu, X.; Zhang, M.; Gu, Q.; Ren, Y.; He, D.; Gao, W. Named Entity Recognition of Fresh Egg Supply Chain Based on BERTCRF Architecture. *Trans. Chin. Soc. Agric. Mach.* **2021**, *52*, 519–525.
4. Yang, Y.; Li, Y.; Zhong, X.; Xu, L. Named Entity Recognition of TCM Medical Records Based on BiLSTM-CRF. *Inf. Tradit. Chin. Med.* **2021**, *38*, 15–21. [CrossRef]
5. Xu, L.; Li, J. Biomedical named entity recognition based on BERT and BiLSTM-CRF. *Comput. Eng. Sci.* **2021**, *43*, 1873–1879.
6. Shen, T.; Yu, L.; Jin, L.; Huang, F.L.; Xv, H.Q. Chinese entity recognition based on BERT-BiLSTM-CRF model. *J. Qiqihar Univ. (Nat. Sci. Ed.)* **2022**, *38*, 26–32.
7. Malarkodi, C.S.; Lex, E.; Devi, S.L. Named Entity Recognition for the Agricultural Domain. *Res. Comput. Sci.* **2016**, *117*, 121–132.
8. Guo, X.; Zhou, H.; Su, J.; Hao, X.; Tang, Z.; Diao, L.; Li, L. Chinese agricultural diseases and pests named entity recognition with multi-scale local context features and self-attention mechanism. *Comput. Electron. Agric.* **2020**, *179*, 105830. [CrossRef]
9. Yan, L. Automatic Question Answering System for Grape Diseases and Pests Based on Knowledge Graph. Master's Thesis, College of Information Engineering Northwest A&F University, Yangling, China, 2021. [CrossRef]
10. Yu, H.; Shen, J.; Bi, C.; Liang, J.; Chen, H. Intelligent diagnostic system for rice diseases and pests based on knowledge graph. *J. South China Agric. Univ.* **2021**, *42*, 105–116.
11. Li, Y. Research on the Construction of Knowledge Graph of Crop Diseases and Pests. Master's Thesis, Agricultural Information Institute Graduate School, Anyang, China, 2021. [CrossRef]
12. Ren, N.; Bao, T.; Shen, G.; Guo, T. Fine-Grained Named Entity Recognition Based on Deep Learning: A Case Study of Tomato Diseases and Pests. *Inf. Sci.* **2021**, *39*, 96–102. [CrossRef]
13. Zheng, Y.; Wu, H.; Zhu, D.; Chen, B.; Li, W. Question and Answer System Based on the Knowledge Graphs of Litchi and Longan Diseases and Insect Pests. *Comput. Digit. Eng.* **2021**, *49*, 2618–2622.
14. Wang, Q.; Liu, Y.; Yang, N. *Identification and Control of Wheat Diseases and Pests*; Ningxia People's Publishing House: Ningxia, China, 2009.
15. Shang, H.; Wang, F. *Atlas of Diagnosis and Control of Wheat Diseases and Pests*; Jindun Publishing House: Beijing, China, 2015.
16. Shen, L.; Jiang, H.; Hu, B.; Xie, C. A study on joint entity recognition and relation extraction for rice diseases pests weeds and drugs. *J. Nanjing Agric. Univ.* **2020**, *43*, 1151–1161.
17. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.
18. Pennington, J.; Socher, R.; Manning, C. Glove: Global Vectors for Word Representation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014.
19. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
20. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.

21. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv* **2019**, arXiv:1909.11942.
22. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
23. Graves, A.; Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.* **2005**, *18*, 602–610. [CrossRef] [PubMed]
24. Lafferty, J.; Mccallum, A.; Pereira, F.C.N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the 18th International Conference on Machine Learning 2001, Williamstown, MA, USA, 28 June–1 July 2001; pp. 282–289.
25. Strubell, E.; Verga, P.; Belanger, D.; McCallum, A. Fast and accurate entity recognition with iterated dilated convolutions. *arXiv* **2017**, arXiv:1702.02098.
26. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF models for sequence tagging. *arXiv* **2015**, arXiv:1508.01991.
27. Dai, Z.; Wang, X.; Ni, P.; Li, Y.; Li, G.; Bai, X. Named entity recognition using BERT BiLSTM CRF for Chinese electronic health records. In Proceedings of the 2019 12th International Congress on Image and Signal Processing, Biomedical Engineering and Informatics (CISP-BMEI), Suzhou, China, 19–21 October 2019; IEEE: Toulouse, France, 2019; pp. 1–5.