agriculture

MDPI

Article

ACE-ADP: Adversarial Contextual Embeddings Based Named Entity Recognition for Agricultural Diseases and Pests

Xuchao Guo [1], Xia Hao [2], Zhan Tang [1], Lei Diao [1], Zhao Bai [1], Shuhan Lu [3] and Lin Li [1,*]

[1] College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China; gxc@cau.edu.cn (X.G.); styx_tang@cau.edu.cn (Z.T.); S20193081368@cau.edu.cn (L.D.); s20193081367@cau.edu.cn (Z.B.)
[2] College of Information Science and Engineering, Shandong Agricultural University, Tai'an 271000, China; haoxia@sdau.edu.cn
[3] School of Information, University of Michigan, Ann Arbor, MI 48104, USA; shuhanlu@umich.edu
* Correspondence: lilinlsl@cau.edu.cn

Abstract: Entity recognition tasks, which aim to utilize the deep learning-based models to identify the agricultural diseases and pests-related nouns such as the names of diseases, pests, and drugs from the texts collected on the internet or input by users, are a fundamental component for agricultural knowledge graph construction and question-answering, which will be implemented as a web application and provide the general public with solutions for agricultural diseases and pest control. Nonetheless, there are still challenges: (1) the polysemous problem needs to be further solved, (2) the quality of the text representation needs to be further enhanced, (3) the performance for rare entities needs to be further improved. We proposed an adversarial contextual embeddings-based model named ACE-ADP for named entity recognition in Chinese agricultural diseases and pests domain (CNER-ADP). First, we enhanced the text representation and overcame the polysemy problem by using the fine-tuned BERT model to generate the contextual character-level embedded representation with the specific knowledge. Second, adversarial training was also introduced to enhance the generalization and robustness in terms of identifying the rare entities. The experimental results showed that our model achieved an F_1 of 98.31% with 4.23% relative improvement compared to the baseline model (i.e., word2vec-based BiLSTM-CRF) on the self-annotated corpus named Chinese named entity recognition dataset for agricultural diseases and pests (AgCNER). Besides, the ablation study and discussion demonstrated that ACE-ADP could not only effectively extract rare entities but also maintain a powerful ability to predict new entities in new datasets with high accuracy. It could be used as a basis for further research on other domain-specific named entity recognition.

Keywords: digital agriculture; Chinese agricultural diseases and pests; named entity recognition; adversarial training; semantic enhancement

Citation: Guo, X.; Hao, X.; Tang, Z.; Diao, L.; Bai, Z.; Lu, S.; Li, L. ACE-ADP: Adversarial Contextual Embeddings Based Named Entity Recognition for Agricultural Diseases and Pests. *Agriculture* **2021**, *11*, 912. https://doi.org/10.3390/agriculture11100912

Academic Editor: Dimitre Dimitrov

Received: 5 September 2021
Accepted: 22 September 2021
Published: 24 September 2021

1. Introduction

Agricultural diseases and pests (ADPs) are one of the major disasters in the world. According to the statistics from the Food and Agriculture Organization of the United Nations (FAO), the global annual economic loss caused by ADPs exceeds US$290 billion [1]. Therefore, how to realize the early detection and early control of ADPs is very important to reduce the losses. With the rapid development of the Internet, agricultural diseases and pests-related text data have shown explosive growth, but it is difficult to be directly recognized and used by computers because of its irregularities and unstructured. The knowledge graph is essentially a semantic web, which can integrate scattered, irregular, and unstructured text data into the agricultural knowledge base. As the basic component of knowledge graph construction and question answering, the named entity recognition task is applied into digital agriculture by some knowledge graph-based human-computer diagnostic systems (e.g., website-based AI question answering systems and diagnostic

systems) to identify the agricultural diseases and pests-related nouns such as "Wheat scab", "Echinocereus squameus", and "Carbendazim" from the texts collected on the Internet or users' inputs on the diagnostic systems so that to extend the agricultural knowledge graph and provide the general public with the solutions for the crop diseases and pests. It has gradually extended from the general field that extracting person and location to specific fields such as geography [2], clinical medicine [3,4], and finance [5]. However, there is still room for improvement to identify the agricultural diseases and pests-related named entities, which has important research value and practical significance for the prevention and control of agricultural diseases and pests and serving modern agriculture.

The purpose of CNER-ADP is to identify the named entities related to agricultural diseases and pests from texts. However, the following limitations in text data and NER models increase the difficulties of recognizing the named entities in agricultural diseases and pests. (1) It is insufficient annotated data in the agricultural domain, and even it is very difficult to collect enough raw text, which also occurs in other domain-specific fields [6]. Taking agriculture as an example, apart from our self-annotated corpus AgCNER [7], there is no publicly available annotated dataset, which directly hinders the research of agricultural named entity recognition. (2) Furthermore, it is impractical to solve the problem of named entity recognition in the field of agricultural diseases and pests with the help of datasets or pre-trained models in other fields, since the texts in different fields usually contain different proper nouns [8,9]. Taking Figure 1 as an example, the agricultural texts contain many domain-specific proper nouns such as "Carbendazim" and "Edifenphos", which are different in semantics from the nouns such as "right hip" and "back hip" in the field of clinical medicine and "wall calendar" and "Ware" in literature [10].

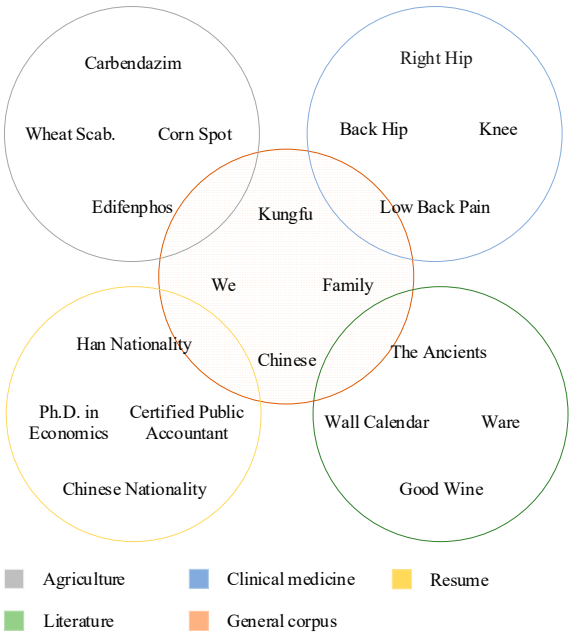


Figure 1. Visualizations of data spaces of the datasets in different fields.

In the case of NER models, as far as we know, the research of named entity recognition in agricultural diseases and pests starts relatively late compared with other domains such as social media and biomedical science [11]. The traditional methods such as rule-based methods, dictionary-based methods, and machine learning-based methods were mainly used to recognize the agricultural diseases and pests named entities [12–14]. The rule- and dictionary-based methods need to pre-design the rules or collect the dictionaries; all of them are less flexible. The machine learning-based methods such as support vector machine (SVM), naive Bayes, and conditional random field (CRF) heavily rely on the manual features, which results in not only the waste of time and effort but also an inability to meet the

requirements of massive and complex texts, which have been reported in many previous works [3,15]. In recent years, deep learning, which can realize end-to-end learning without using hand-designed features, has brought breakthroughs for computer vision fields such as image classification [16–18], semantic segmentation [19], and object detection [20], and natural language processing such as text classification, machine translation, and knowledge question answering. However, there only a few works have begun to consider it, recognizing the agricultural named entities [7,15,21]. The common problem of the above models is that all of them utilize the traditional word embedding methods (e.g., word2vec [22]) to generate context-independent embeddings, which cannot effectively solve the polysemy problem, i.e., the same word may have different meanings in different contexts. For example, "Kung Fu" refers to a sport in the general field, while in the field of agricultural diseases and pests, and it refers to the name of a drug. In addition, word2vec can only learn the shallow semantic features, but it is limited in the extraction of syntax, semantics, and other high-level features [23]. In the previous work [24], the pre-trained model, such as the bidirectional encoder representation from transformers (BERT), was used to generate the context-sensitive embeddings, and the CRF was also considered as the decoder to predict the final labels. Due to the difference in data distribution between the domain-specific texts, the original BERT may be limited in the representation of specific knowledge. Recent studies have shown that there is a certain proportion of rare entities in agricultural texts, and the performance of most existing models for such entities needs to be further improved [7].

1.1. Recent Developments Related to NER Models

Researchers try to improve the models' performance mainly from two aspects, i.e., contextual encoders and text representation. Most models try their best to improve the ability to capture the useful text representation by designing efficient neural network architectures, including the commonly used contextual encoders such as Convolutional Neural Networks (CNN), Bi-directional Long Short-Term Memory (BiLSTM) [25], and their variants [26–31]. The former has significant advantages in extracting global context features, and the latter is good at capturing local context features, which are as useful as global context features. Some studies integrated the above two architectures and then proposed hybrid models, e.g., CNN-BiLSTM-CRF, to make full use of the two types of context features [7]. Other works integrated the self-attention mechanism to enhance the ability to captured long-term dependencies [3,32,33]. Moreover, some other typical variants of CNN, such as Gated CNN [27,34], RD_CNN [28], GRN [30], and CAN [31], were also proposed to recognize the entities. Besides, the transformer-based models (e.g., TENER [29] and FLAT [35]) and graph neural network-based models [36,37] have gradually attracted considerable attention in recent years. However, high-quality text representation is the prerequisite and basis for the improvement of the overall performance of the NER models; that is, text representation should contain as much knowledge as possible, such as syntax, semantics, word meaning, and so on. Otherwise, even if the context encoder maintains a strong ability of feature extraction, it may not significantly improve the final recognition accuracy [38]. There is another challenge, i.e., existing models cannot effectively identify the rare words in agricultural texts [8].

Text representation is an effective method that describes the text features by converting discrete text sequences into low-dimensional dense vectors [39]. In the early stage, non-contextual embeddings models were often used to learn shallow semantic features. Some works utilized word2vec or glove to pre-train the lookup table of word embeddings and applied it into named entity recognition [3,40]. Until now, the non-contextual embeddings models are still used to generate the word-level or character-level embeddings [41,42]. Xin Liu et al. [43] introduced a deep neural network, named OMINer, for online medical entity recognition; they also pre-trained the word2vec on a large-scale corpus to produce a lookup table that can be used for Chinese online medicine query text. Besides, some works attempt to use CNN and BiLSTM to further extract and integrate external knowledge such as radical

and morphological features [3,44,45]. Although the performance of the NER model is slightly improved, word2vec has obvious limitations, i.e., they fail to distinguish different semantic information of the polysemous words and cannot extract high-level features such as syntactic structure. Recently, the language models have brought a milestone breakthrough for many natural language processing tasks. However, the available BERT model for Chinese was pre-trained on the Chinese Wikipedia corpus, which belongs to the general field. It is undeniable that the pre-trained BERT performs well in the general domain but is not efficient in specific fields [9]. Furthermore, due to the limited corpus in agricultural domains, it is unable to provide enough data for pre-training. Fine-tuning is a commonly used compromise method, which can not only solve the problem of limited data but also help the language model to learn the knowledge of specific fields [46]. Different from English and other Latin characters, Chinese characters are hieroglyphs, and their morphological structure contains rich glyph features, which can be extracted from the perspective of images and are helpful to Chinese named entity recognition. For example, Song and Sehanobish [47] managed to integrate the fine-tuned BERT and extract the glyph features for Chinese NER. Based on the above work, Xuan et al. [48] proposed a fusion glyph network to further explore the interaction between the glyph features and the contextual embeddings. In short, fine-tuning can improve the quality of text representation in the case of lacking data. However, because the BERT is task-agnostic, the limited training dataset may not cover all the semantic features in the field, which will affect the overall robustness and generalization of the NER models.

1.2. Objectives and Hypotheses

To address the abovementioned issues, a general method for agricultural diseases and pests named entity recognition, named ACE-ADP, was proposed in this paper. The objective of ACE-ADP was to use the pre-trained language model (i.e., BERT, which would be fine-tuned on the agricultural training dataset) to learn the domain-specific features. The text representation would be enhanced by the fine-tuned BERT with agricultural knowledge. Besides, adversarial training would also be introduced to enhance robustness and generalization in terms of identifying rare entities. In the course of this study, the following hypotheses were tested:

(1) An adversarial contextual embeddings-based model could be applied for agricultural diseases and pests named entity recognition. As far as we know, it was the first time that combined BERT and adversarial training to recognizing the named entities in the field of agricultural diseases and pests;
(2) The BERT, which was fine-tuned on the agricultural corpus, could generate the high-quality text representation so that to enhance the quality of text representation and solve the polysemous problem;
(3) Adversarial training could also be adopted to solve the rare entity recognition problem. Besides, it could also exert its maximum performance when the text representation was of high quality. As far as we know, the previous research had not explicitly raised this point;
(4) ACE-ADP could significantly improve the F_1 of CNER-ADP with an improvement of 4.31%, especially for rare entities, in which an F_1 was increased by 9.83% on average.

We organized the rest of the paper as follows. The experimental corpora, parameter settings, evaluation metrics, and the proposed method were introduced in Section 2. The experimental results and ablation study are presented in Section 3. The discussions are conducted in Section 4. The conclusion and future directions are described in Section 5.

2. Materials and Methods

We implemented the ACE-ADP with the TensorFlow framework and ran on a single GTX 1080 Ti GPU, Windows 10. The source code will be released at https://github.com/guojson/ACE-ADP.git (accessed date: 15 September 2021).

2.1. Datasets

We assessed the performance of our proposed method on four benchmark datasets, i.e., AgCNER [7], CLUENER [49], CCKS2017, and Resume [50]. Among them, AgCNER is an agricultural dataset that was annotated by ourselves in previous works [7] and includes 11 types of entities related to agricultural diseases and pests. The data distribution for each category in AgCNER is illustrated in Figure 2a. It can be seen from the figure that in addition to a large number of entities such as crop, disease, and pest, there are also some rare entities such as pathogeny, weed, and fertilizer, which undoubtedly increases the difficulty of CNER-ADP task. CLUENER is collected from THUCNews and contains 10 fine-grained entity categories such as Finance and Stock. CCKS2017, released by the 2017 China Conference on Knowledge Graph and Semantic Computing, contains five clinical medicine-related categories and 2231 annotated samples. According to Figure 2b, its data distribution for each category is relatively balanced, but there are also some difficulty-to-identify categories address, scene, and book need to be further considered [49]. The details of all datasets were listed in Table 1. Note that all datasets were labeled by BIO scheme (i.e., Begin, Inside, and Other) and divided into the training set and test set according to the ratio of 8:2. For word2vec-based models, we exploited the character-level embeddings that pre-trained on the Baidu Baike corpus [51].

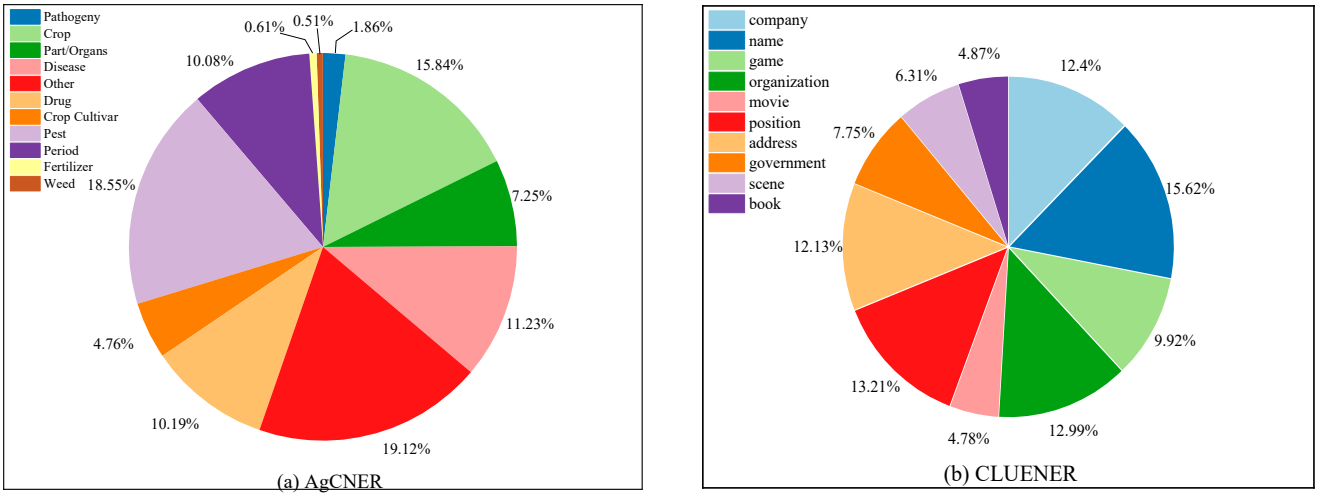


Figure 2. The proportion of each category in the data sets. (**a**) Illustrates the data distribution of each category in AgCNER; (**b**) shows the data distribution of each category in CLUENER.

Table 1. The detailed information of all datasets.

Dataset	Domain	Samples	Entities	Class	Categories
AgCNER	Agriculture	24,696	248,171	11	Crop, Disease, Drug, Fertilizer, Part/Organs, Period, Pest, Pathogeny, Crop Cultivar, Weed, Other
CLUENER	News	12,091	26,320	10	Person, Organization, Position, Company, Address, Game, Government, Scene, Book, Movie
CCKS2017	Clinic	2231	63,063	5	Body, Symptoms, Check, Disease, Treatment
Resume	Resume	4740	16,565	8	Country, Educational institution, Location, Personal name, Organization, Profession, Ethnicity, Background and Job, Title

2.2. Parameter Setting

During the training process, the exponential decay function was used to dynamically control the learning rate and thus to control the speed of parameter updating. In this paper, the decay rate was set to 0.9, and the decay step was 5000. The learning rate for BERT was

set to 5×10^{-5} and 0.0001 for the NER model during the fine-tuning process. Moreover, it was set to 0.002 on AgCNER and 0.001 on other data sets during the training process. In this paper, early stopping [52] and a patience of 10 was used to prevent the over-fitting problem. Other hyper-parameters are listed in Table 2.

Table 2. Parameter settings for ACE-ADP model.

Hyper-Parameter		Value
Character embedding		768
Hidden units		256
Dropout		0.25
Optimizer		Adam
Batch_size	fine-tuning	8
	model training	32
Max_epoch	Word2vec	100
	BERT	50

2.3. Evaluation Metrics

In this paper, *Precision* (*P*), Recall (*R*), and F_1-*score* (F_1) were used as the evaluation metrics; only the boundary and type were both correctly identified, the entity could be correctly predicted. Note that their units were %, the below was the same. We ran the experiment three times according to [8], and the average results with standard deviation were listed:

$$Precision = \frac{TP}{TP + FP}, \tag{1}$$

$$Recall = \frac{TP}{TP + FN}, \tag{2}$$

$$F_1\text{-}score = \left(1 + \frac{FP + FN}{2TP}\right)^{-1}. \tag{3}$$

where TP represents the number of labels that are positive and predicted to be positive. FP represents the number of labels that are negative and predicted to be positive. FN represents the number of labels that are negative and predicted to be negative.

2.4. ACE-ADP Method

2.4.1. Problem Definition

In this paper, we regard the named entity recognition task as the sequence labeling problem. Given a sentence $S = (c_1, c_2, \ldots, c_n,)$ with length n, where c_i represents the i-th Chinese character. Generally speaking, the discrete sentence S will be converted into low-dimensional dense embeddings, i.e., $E = (e_1, e_2, \ldots, e_n,)$, where $e_i \in R^d$ donates the embedding vector of c_i. Then E will be fed into context encoders (e.g., BiLSTM or CNN) to extract the context features. Next, the decoder (e.g., softmax and CRF) will be exploited to predict the gold label $\hat{y}_i$ (e.g., B-LOC, I-LOC, and O) for each character c_i. Finally, the predicted labels $\hat{Y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_i)$ for sentence S to be obtained, and the entities maintained in a sentence will be recognized. Formally, the object of the NER is to learn a function $f_\theta : S \rightarrow \hat{Y}$ to predict the labels for all characters.

2.4.2. Fine-Tuned BERT

As described in Section 1, obtaining high-quality embeddings is the first step for the NER model to predict the labels. Different from the early works that utilized word2vec to generate the context-independent embeddings, in this paper, BERT was considered as a generator to produce the context-sensitive embeddings according to the different contexts. BERT is composed of N layers of bidirectional Transformer blocks, which is more efficient to capture the deeper bidirectional relationships by jointly modeling the forward and

backward contexts of each word. Formally, we define transformer blocks as $Trans(x)$, then the embedding vector E will be obtained as follows:

$$E_0 = S'W_e + W_p, \tag{4}$$

$$E_l = Trans(E_{l-1}),\ l \in [1,\ N], \tag{5}$$

where S' is the one-hot matrix corresponding to sentence S, W_e represents the embedding matrix pre-trained by BERT, W_p donates the positional embeddings that can be calculated by Equations (6) and (7). E_l represents the contextual embedding at the l-th layer. N is the number of layers of transformer blocks. In this paper, N was set to 12.

$$W_{(p_i,2i)} = \sin\left(p_i/10000^{2i/d}\right), \tag{6}$$

$$W_{(p_i,2i+1)} = \cos\left(p_i/10000^{2i/d}\right), \tag{7}$$

In terms of CNER-ADP, there is a general lack of corpus, which cannot provide sufficient data support for the pre-training of BERT. In this paper, fine-tuning was regarded as a compromise solution to alleviate the insufficient corpus to a certain extent. First of all, BERT parameters were initialized by using the original weights pre-trained on the Chinese Wikipedia corpus (https://storage.googleapis.com/bert_models/2018_11_03/chinese_L-12_H-768_A-12.zip), which belongs to the general domain. Then, a fully connected network was used on the top layer of the BERT to obtained the 768-dimensional context representation. Different from the BERT-CRF architecture proposed in [48], the context encoder, i.e., BiLSTM was integrated between the BERT and CRF to further extract global context features. The fine-tuning architecture for CNER-ADP was shown in Figure 3 without the component of adversarial perturbation. Besides, the fine-tuned weights were saved separately to initialize another BERT used in the models of CNER-ADP, for the reason that the learning rate for fine-tuning is minimal while the training requires a larger one. Moreover, freeze BERT contributes to decreasing the computation and storage load, which is also an important factor to be considered.

2.4.3. Context Encoder and Decoder

In this paper, BiLSTM was used as the encoder to further extract the contextual features from the text representation. It is a variant of RNN and can efficiently solve the problems of gradient vanishing and gradient explode. The formal description for a single LSTM cell is shown in Equations (8)–(10):

$$\begin{bmatrix} f_t \\ i_t \\ o_t \\ \widetilde{C}_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ tanh \end{bmatrix} \left(\begin{bmatrix} W \\ U \end{bmatrix}^T \begin{bmatrix} e_t \\ h_{t-1} \end{bmatrix} + b \right), \tag{8}$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t, \tag{9}$$

$$h_t = o_t * \tan \mathrm{h}(C_t), \tag{10}$$

where f_t, i_t, o_t, $\widetilde{C}_t$, and C_t represent the forget gate, input gate, output gate, candidate cell state, and the memory state at time step t, respectively. The sigmoid is used as the activation function $\sigma(*)$, W, U, and b are trainable parameters. e_t is input at time step t and h_{t-1} is the hidden state at the last timestep. At each time step t, BiLSTM will generate forward and backward hidden vectors $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$, which maintain the forward and backward context information, respectively. The output of BiLSTM at timestep t will be obtained, i.e., $h_t = \left[\overrightarrow{h_t}; \overleftarrow{h_t}\right]$ with dimension $2d_c$ and the final output for sentence S is defined as $H = (h_1, h_2, \ldots, h_n)$.

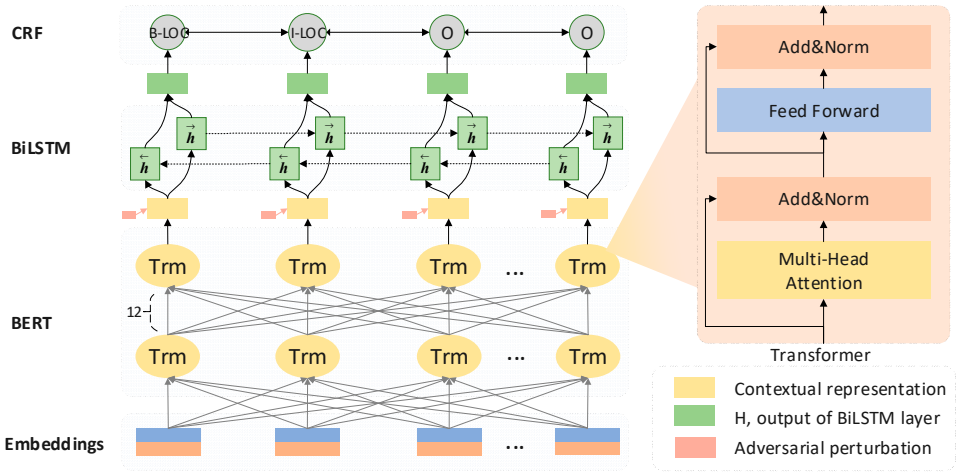


Figure 3. The architecture of the ACE-ADP model for agricultural diseases and pests.

In this paper, the CRF was considered as the decoder because of the strong dependency between adjacent labels in sequence labeling tasks. For example, B-LOC is usually followed by I-LOC but cannot be followed by B-PER or I-PER. Therefore, joint decoding may be more beneficial than independent decoding [53]. As shown in Equation (11), given the predicted tags $\hat{Y}$ of sentence S and its corresponding embedding vector E, its score is calculated by the state score $P \in R^{n \times d_r}$ and state transition matrix $T \in R^{d_r \times d_r}$. Among them, P is mapped from H by a fully connected layer (Equation (12)). Thus, the probability of predicted labels $\hat{Y}$ in all possible tag sequences Y_{all} is calculated by Equation (13):

$$Score(\hat{Y}, S) = \sum_{i=0}^{n} P_{i,\,\hat{y}_i} + \sum_{i=0}^{n-1} T_{\hat{y}_i, \hat{y}_{i+1}}, \tag{11}$$

$$P = HW_p + b_p, \tag{12}$$

$$p(\hat{Y}|S) = \frac{\exp\left(Score(S, \hat{Y})\right)}{\sum_{\hat{Y}' \in Y_{all}} \exp\left(Score(S, \hat{Y}')\right)}, \tag{13}$$

where $W_p \in R^{2d_c \times d_r}$ and $b_p \in R^{n \times d_r}$ are trainable parameters, d_r is the number of tags. The gold tag sequence with the highest score is obtained by the Viterbi algorithm.

2.4.4. Adversarial Training

To solve the over-fitting problem and enhance the ability to recognize the rare entities, we treated adversarial training as a data augmentation method, i.e., a new adversarial sample would be generated after adding a small perturbation to the training sample. Assuming that the loss function of the ACE-ADP model without adversarial training was shown in Equation (14). Y represents the ground truth labels. The goal of our model is to minimize the loss by training the weights θ.

$$loss(\hat{Y}, Y) = -\sum logp(\hat{y}|E; \theta), \tag{14}$$

$$loss(\hat{Y}_{adv}, Y) = -\sum logp(\hat{y}|E + \eta_{adv}; \theta), \tag{15}$$

As shown in Equation (15), the adversarial training guided loss function would be obtained after adding a worst-case perturbation η_{adv} to the embeddings. In general, η_{adv} can be calculated by the following function:

$$\eta_{adv} = \underset{\eta, ||\eta|| \leq \varepsilon}{\text{argmin}} logp(\hat{y}|E + \eta; \hat{\theta}), \tag{16}$$

where η is a perturbation, ε is the bounded norm, which can be calculated by $\varepsilon = \gamma\sqrt{d}$ according to [8], d is the dimension of embeddings, γ is perturbation size that should be

reasonably selected for the reason that if γ is too small to play the role of perturbation. Conversely, it will easily introduce noise that can destroy the original semantic information. $\hat{\theta}$ presents the current training weights of the model. Due to the non-differentiability of Equation (16), similar to [54], the approximation is used to replace η_{adv}, as shown in Equation (17).

$$\eta_{adv} = -\frac{\varepsilon g}{\| g \|_2}, \text{where} g = \nabla_E log p\left(\hat{y} \middle| E; \hat{\theta}\right), \tag{17}$$

The loss function of the model with adversarial training was defined as follows in Equation (18).

$$loss = loss\left(\hat{Y}, Y\right) + loss\left(\hat{Y}_{adv}, Y\right), \tag{18}$$

The steps of our proposed model can be found in Appendix A. Based on the above descriptions, the advantages of innovative works for the CNER-ADP task can be summarized as follows:

(1) Contextual-sensitive. BERT can dynamically generate the context-dependent embeddings according to the contexts, which is beneficial for solving the problem of polysemous words that are often caused by context-independent methods such as word2vec and glove;
(2) Domain-aware. In this paper, domain knowledge can be injected into BERT by fine-tuning, which is essential to handle the NER task in specific domains;
(3) Stronger robustness and generalization. The experimental results in Section 4.4 showed that compared with previous models, our proposed model maintains high robustness and generalization.

3. Results

To verify the effectiveness of the ACE-ADP, we conducted comprehensive experiments with several state-of-the-art models on the self-annotated agricultural datasets AgCNER and three other corpora that belong to different fields. The experimental results showed that the proposed model achieved remarkable results and could significantly improve the accuracy of difficult-to-identify entities such as the entities with fuzzy boundaries and the rare ones.

3.1. Main Results Compared with Other Models

The results of all models (i.e., ACE-ADP and several state-of-the-art models proposed in recent years) on four datasets were listed in Table 3. Note that we exploited the fine-tuning BERT for IDCNN, Gated CNN, and AR-CCNER [3] to obtain the best results, and others were set according to their original papers. Our proposed model achieved the highest F_1 of 93.68%, 98.31%, 95.72%, and 96.83% on CLUENER, AgCNER, CCKS2017, and Resume, respectively. For example, ACE-ADP outperformed the IDCNN with improvements of 3.57%, 15.7% in terms of F_1 on AgCNER and CLUENER due to the effectiveness of adversarial training. Moreover, compared with IDCNN, Gated CNN tended to achieve slightly better F_1 on AgCNER and Resume, which benefits from the gated structure that can filter useful features according to their importance. However, due to the blurring boundaries of entities in CLUENER and ccks2017, it performed slightly worse than IDCNN. In contrast, AR-CCNER has achieved a few higher F_1-*scores* than IDCNN and Gated CNN on most datasets, thanks to the fact that radical features may provide rich external knowledge, and the self-attention mechanism helps to enhance the model's ability to capture the long-distance dependencies.

Moreover, we also conducted experiments with other state-of-the-art models, i.e., FGN [48], TENER [29], and Flat-Lattice [35]. FGN not only outperformed IDCNN, Gated CNN as reported by Xuan et al. [48], who integrated the interactive information between the contextual embeddings generated by the fine-tuning BERT and the glyph information extracted by novel CNN structure, but also better than TENER, a transformer-based model, indicating that the fine-tuning BERT may outperform a single transformer-based model in

NER task. Moreover, Flat-Lattice, which benefits from the flat-lattice Transformer and the well-designed position encoding, also presented remarkable results. However, the number of potential words will be increased as the length of the sentence increases, which tends to result in a significant increase in the structural complexity. Unlike FGN and Flat-Lattice, apart from the basic framework BiLSTM-CRF, our model only utilized the fine-tuned BERT and adversarial training to enhance the robustness and generalization, showing lower structural complexity. The results of ACE-ADP went beyond previous reports and slightly lower standard deviations showing its positive effect on domain-specific (e.g., the agricultural diseases and pests) named entity recognition task.

Table 3. Experimental results for all models on four different datasets.

Algorithms	CLUENER			AgCNER			CCKS2017			Resume		
	P	***R***	F_1	***P***	***R***	F_1	***P***	***R***	F_1	***P***	***R***	F_1
BERT-IDCNN-CRF	78.37	77.60	77.98 ± 0.11	94.39	95.08	94.74 ± 0.07	90.55	93.52	92.01 ± 0.13	95.47	96.59	96.03 ± 0.13
BERT-Gated CNN-CRF	75.85	77.98	76.90 ± 0.34	94.32	95.20	94.76 ± 0.08	89.43	92.93	91.15 ± 0.14	95.75	96.81	96.27 ± 0.16
AR-CCNER	78.34	77.74	78.04 ± 0.28	94.60	94.73	94.67 ± 0.06	90.23	93.36	91.77 ± 0.28	95.89	97.22	96.55 ± 0.27
FGN [48]	79.50	79.71	79.60 ± 0.15	94.33	94.56	94.45 ± 0.03	90.44	93.09	91.75 ± 0.16	96.67	97.09	96.88 ± 0.10
TENER	72.94	74.21	73.57 ± 0.17	93.01	95.22	94.10 ± 0.09	91.24	93.08	92.15 ± 0.13	94.91	95.03	94.97 ± 0.21
Flat-Lattice [35]	79.25	80.68	79.96 ± 0.13	93.52	94.31	93.91 ± 0.08	91.55	93.40	92.46 ± 0.16	95.22	95.72	95.47 ± 0.18
ACE-ADP	**93.03**	**94.36**	**93.68 ± 0.18**	**98.30**	**98.32**	**98.31 ± 0.02**	**95.17**	**96.27**	**95.72 ± 0.13**	**96.22**	**97.44**	**96.83 ± 0.17**

3.2. Ablation Study

3.2.1. Macro-Level Analysis

The contextual embeddings based on fine-tuned BERT and adversarial training were the focus of this paper. An ablation study was first conducted to verify their effectiveness and necessity from a macro-level perspective. The experimental results are listed in Table 4. Taking the AgCNER and Resume as an example, according to groups 1 and 4, the model with adversarial training improves the F_1 on AgCNER and Resume by 3.43% and 0.97%, respectively, indicating that adversarial training helps to improve the performance of named entity recognition of the models. Besides, the F_1-*scores* in groups 1, 2, and 5 presented that adversarial training was positively related to the quality of text embeddings, i.e., the higher the quality of text representation, the better the effect of adversarial training. From the results of groups 2, 3, and 1, we could observe that the word2vec-based model (group 2) showed the worst performance on all datasets, the possible reason is that the text embeddings generated by word2vec are context-independent, which cannot provide enough semantic information for the model training. In contrast, the models that integrated the original and fine-tuned BERT delivered significantly better F_1-*scores*, i.e., 96.11% and 98.31% on AgCNER, and 96.38% and 96.83% on Resume respectively, due to the high-quality context embeddings based on BERT in the case of adversarial training. Meanwhile, the fine-tuned BERT-based model (group 1) presented better performance than the original BERT-based model (group 3) for the reason that the contextual embeddings generated by BERT express abundant semantic information, which has a positive effect on improving the performance of the model. Moreover, fine-tuning enables BERT to obtain domain awareness and makes the contextual embeddings contain more domain-specific knowledge, which is crucial for domain-specific NER tasks. In particular, compared with the baseline model listed in group 5 (i.e., word2vec-based BiLSTM-CRF), the F_1-*score* of the proposed model on the AgCNER was increased by 4.23%. Similar results could also be presented on other datasets. In short, the presented findings confirmed the necessity and effectiveness of contextual embeddings and adversarial training.

Table 4. Recognition results of ACE-ADP and its variants on four datasets.

#	Algorithms	CLUENER			AgCNER			CCKS2017			Resume		
		P	*R*	F_1	*P*	*R*	F_1	*P*	*R*	F_1	*P*	*R*	F_1
1	ACE-ADP	93.03	94.36	93.68 ± 0.18	98.30	98.32	98.31 ± 0.02	95.17	96.27	95.72 ± 0.13	96.22	97.44	96.83 ± 0.17
2	-BERT	68.43	67.15	67.78 ± 0.29	94.01	93.89	93.95 ± 0.06	90.27	91.86	91.05 ± 0.23	91.25	93.15	92.19 ± 0.15
3	-fine-tuning	92.02	93.16	92.58 ± 0.13	95.99	96.23	96.11 ± 0.17	95.01	97.15	96.07 ± 0.16	95.78	96.85	96.38 ± 0.09
4	-AT	78.83	77.39	78.11 ± 0.02	94.59	95.16	94.88 ± 0.04	90.30	92.84	91.56 ± 0.14	95.12	96.60	95.86 ± 0.28
5	-BERT-AT	68.48	66.95	67.70 ± 0.41	94.18	93.99	94.08 ± 0.06	89.16	91.42	90.27 ± 0.13	92.09	93.56	92.82 ± 0.12

'-' means not participating in training.

3.2.2. Effect of BERT

To further verify the effectiveness of BERT in detail, several experiments with word2vec-, original BERT-, and fine-tuned BERT-based models were conducted on four benchmark datasets. Their F_1-*scores* are presented in Table 5. As expected, the word2vec-based models tended toward the lower F_1-*scores* than BERT-based ones on all datasets, as discussed in Section 3.2.1. BERT, which consists of 12-layer of bidirectional Transformers, could dynamically generate high-quality embeddings according to the different contexts. For example, BiLSTM with original BERT achieved significant improvement of F_1-*scores* with +9.07%, +0.11%, +1.35%, and +2.06% on CLUENER, AgCNER, CCKS2017, and Resume, respectively. However, there was still room for improvement because of the task independence of the original BERT. The fine-tuning BERT-based models have presented the best performance in multiple domain-specific datasets for the reason that BERT not only maintains the strong ability of semantic representation but also obtains the domain awareness after fine-tuning, which may encourage BERT to represent the domain-specific features efficiently [9]. Besides, in the case of BERT, the performance of IDCNN, Gated CNN, AR-CNER, and CNN-BiLSTM-CRF were also improved. Therefore, the present findings demonstrated the effectiveness of BERT, and it would be more suitable for the domain-specific NER tasks after fine-tuning.

3.2.3. Effect of Adversarial Training

The F_1 of the adversarial training-based models with word2vec, original BERT, and fine-tuning BERT were presented in Table 6. Combining with the details presented in Table 5, several important conclusions could be summarized: (1) The word2vec-based model with adversarial training tended towards slightly worse results, indicating that in the case of poor text representation, adding perturbation would be counterproductive. (2) The F_1 of original BERT-based models with adversarial training were significantly improved compared with those listed in Table 5, indicating the effectiveness of the adversarial training to enhance the robustness and generalization. Taking BiLSTM as an example, its F_1 increased by +1.92% on AgCNER and +1.5% on Resume. (3) The F_1 of BiLSTM with fine-tuning BERT and adversarial training were further increased by +2.2% on AgCNER and +0.45% on Resume, which indicated that in the case of original BERT and fine-tuned BERT, the recognition performance could be further improved by using adversarial training. (4) There was very little difference in terms of F_1-*scores* between the BiLSTM, AR-CCNER, and CNN-BiLSTM-CRF, which indicated that the complex architectures such as radical features, self-attention, and CNN might be unnecessary. (5) The experimental results in Tables 5 and 6 show that Gated CNN and RD_CNN achieved better performance than BiLSTM. In actual uses, they could replace BiLSTM as feature encoders to extract local and global context features when integrating high-quality text representation and adversarial training. Furthermore, most of the standard deviations listed in Table 6 are lower than those in Table 5, indicating that the adversarial training may contribute to improving the stability of the model. In short, the above experimental results verified the effectiveness of adversarial training and once again demonstrated that adversarial training could enhance the robustness of the NER model.

Table 5. F_1 of models with word2vec, original BERT, and fine-tuning BERT without adversarial training.

Algorithms	CLUENER			AgCNER			CCKS2017			Resume		
	W	O	F	W	O	F	W	O	F	W	O	F
BiLSTM	67.70 ±0.41	76.77 ±0.35	78.11 ±0.18	94.08 ±0.06	94.19 ±0.07	94.88 ±0.02	90.27 ±0.13	91.62 ±0.16	91.56 ±0.13	92.82 ±0.12	94.88 ±0.11	95.86 ±0.17
IDCNN	66.58 ±0.38	76.33 ±0.23	77.98 ±0.11	93.99 ±0.06	93.91 ±0.13	94.74 ±0.07	91.46 ±0.29	91.20 ±0.31	92.01 ±0.13	92.71 ±0.42	94.44 ±0.33	96.03 ±0.13
Gated CNN	66.26 ±0.25	75.23 +0.22	76.90 ±0.34	93.56 ±0.11	93.72 ±0.02	94.76 ±0.08	91.02 ±0.28	89.86 ±0.12	91.15 ±0.14	89.25 ±0.35	93.12 ±0.23	96.27 ±0.16
RD_CNN	66.16 ±0.15	75.73 ±0.21	77.95 ±0.18	93.20 ±0.08	93.89 ±0.04	94.82 ±0.05	89.18 ±0.23	90.03 ±0.19	91.52 ±0.17	89.56 ±0.23	93.39 ±0.17	95.87 ±0.19
AR-CCNER	68.67 ±0.35	77.08 ±0.26	78.04 ±0.28	94.46 ±0.08	94.12 ±0.06	94.67 ±0.06	91.45 ±0.30	91.10 ±0.15	91.77 ±0.28	93.09 ±0.25	95.01 ±0.19	96.55 ±0.27
CNN-BiLSTM-CRF	68.45 ±0.37	76.88 ±0.22	78.18 ±0.12	94.07 ±0.12	94.53 ±0.02	94.78 ±0.05	92.03 ±0.16	91.49 ±0.24	91.28 ±0.25	93.84 ±0.18	95.18 ±0.16	95.26 ±0.24

"W" represents word2vec, "O" means the original BERT, and "F" donates the fine-tuned BERT.

Table 6. F_1 of adversarial training-based models with word2vec, original BERT, and fine-tuning BERT.

Algorithms	CLUENER			AgCNER			CCKS2017			Resume		
	W	O	F	W	O	F	W	O	F	W	O	F
BiLSTM	67.78 ±0.29	92.58 ±0.13	93.68 ±0.18	93.95 ±0.06	96.11 ±0.17	98.31 ±0.02	91.05 ±0.23	96.07 ±0.16	95.72 ±0.13	92.19 ±0.15	96.38 ±0.09	96.83 ±0.17
IDCNN	66.12 ±0.25	94.72 ±0.21	94.45 ±0.17	93.71 ±0.08	96.98 ±0.14	98.23 ±0.05	91.27 ±0.19	96.12 ±0.13	95.25 ±0.17	93.13 ±0.12	96.91 ±0.11	96.16 ±0.12
Gated CNN	66.07 ±0.14	95.03 ±0.16	96.33 ±0.13	93.48 ±0.03	97.48 ±0.15	98.42 ±0.08	90.88 ±0.12	96.27 ±0.11	96.19 ±0.14	91.57 ±0.16	96.73 ±0.11	97.57 ±0.15
RD_CNN	65.88 ±0.16	94.51 ±0.18	96.68 ±0.13	92.86 ±0.07	97.35 ±0.14	98.95 ±0.05	90.18 ±0.16	95.56 ±0.10	95.61 ±0.15	91.20 ±0.17	96.01 ±0.13	97.34 ±0.14
AR-CCNER	62.30 ±0.36	91.64 ±0.24	89.66 ±0.25	92.80 ±0.11	97.50 ±0.12	97.70 ±0.06	90.96 ±0.20	96.08 ±0.16	95.97 ±0.12	90.36 ±0.15	96.74 ±0.12	97.26 ±0.14
CNN-BiLSTM-CRF	67.23 ±0.26	90.81 ±0.19	89.82 ±0.22	93.67 ±0.11	96.57 ±0.12	97.66 ±0.12	91.63 ±0.16	95.33 ±0.19	95.14 ±0.17	93.07 ±0.16	96.77 ±0.13	96.91 ±0.16

"W" represents word2vec, "O" means the original BERT, and "F" donates the fine-tuned BERT.

4. Discussion

4.1. Performance for Rare Entities

In this section, AgCNER and CLUENER were selected as comparable datasets to verify the performance of ACE-ADP in identifying the rare entities, which are challenging for NER models.

The experimental results were illustrated in Figures 4 and 5, which showed that ACE-ADP outperformed other comparable models and significantly improved the F_1-*scores* of all categories both in AgCNER and CLUENER, especially the categories that are difficult-to-identify or have a low percentage of entities. In terms of AgCNER, ACE-ADP still maintained the highest F_1 on easy-to-identify entities such as disease, pest, and crop, which were 98.98%, 98.73%, and 98.88%, respectively. Meanwhile, it significantly improved the F_1 of rare entities such as fertilizer, weed, and pathogeny by +11.99%, +3.95%, and +8.25% (8.06% on average) compared with the word2vec-based BiLSTM. Moreover, according to the results of CLUENER reported in [49], ACE-ADP could effectively improve the Precision, Recall, and F_1 of difficult-to-recognize entities such as address, scene, and book, all of them achieved F_1-*scores* above 90%. The possible reasons are that adversarial training is helpful to improve the robustness of the model, and fine-tuned BERT can generate the character-level embeddings with rich domain-specific semantic information, which also contributes to improving the NER performance.

In addition, we took the Ft-BERT-BiLSTM and ACE-ADP as examples and visualized their confusion matrices on the AgCNER to further illustrate the effectiveness of the proposed model. As shown in Figure 6, ACE-ADP obtained more correctly predicted labels of the rare entities (e.g., fertilizer, weed, and pathogeny) than Ft-BERT-BiLSTM, which means that ACE-ADP can achieve higher TP, while FP and FN are relatively low.

According to Equation (3), ACE-ADP tends to obtain a higher F_1. Thus, the experimental results illustrated that the high-quality text representation and adversarial training could effectively enhance the NER models' robustness and were useful to identify the rare and difficulty-to-identify entities.

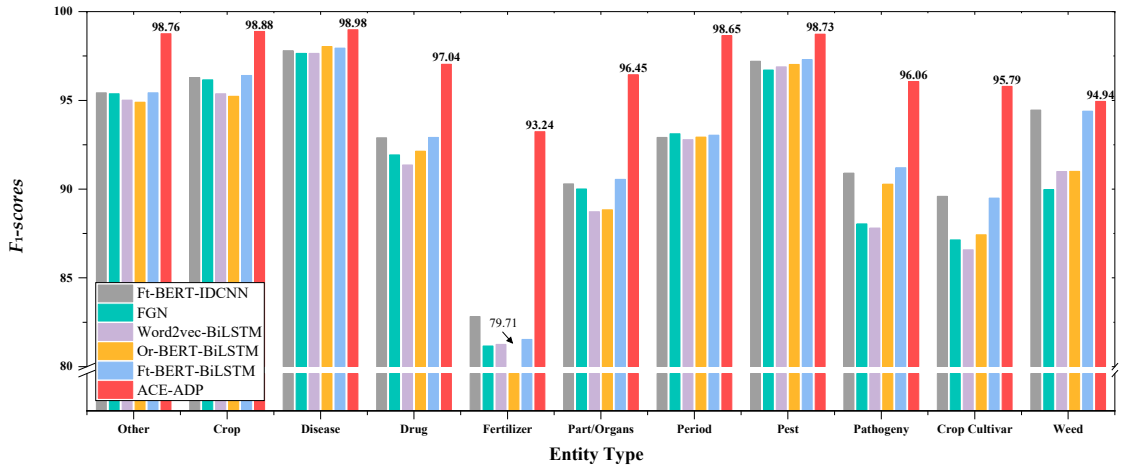


Figure 4. Detailed results of F_1-*scores* for each category on AgCNER. "Ft" means the fine-tuned BERT, "Or" describes the original BERT.

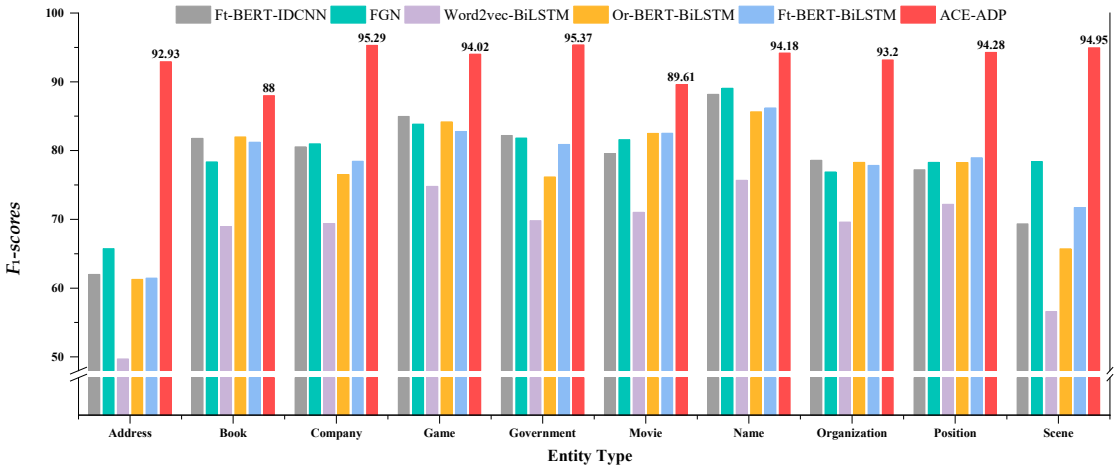


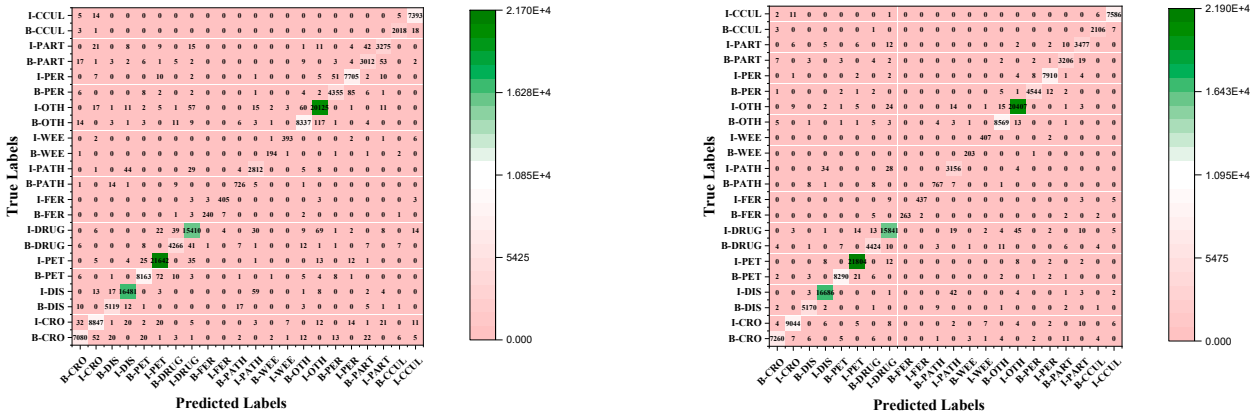
Figure 5. Detailed results of F_1-*scores* for each category on CLUENER.

4.2. Robustness and Generalization

The curves of training and validation loss for Word2vec-BiLSTM-CRF, BERT-BiLSTM-CRF, and ACE-ADP on AgCNER were visualized as Figure 7. As shown in Figure 7a,b, the validation losses of Word2vec-BiLSTM-CRF and BERT-BiLSTM-CRF decrease with the increase in iteration and then gradually increase, showing the obvious over-fitting characteristics, while that of ACE-ADP decreases first and then tends to be flat, indicating that ACE-ADP could effectively alleviate the over-fitting problem.

Apart from the training set and testing set of AgCNER used in this paper, another dataset, which has never been used before and contains 2223 agricultural samples, was considered as the final testing dataset to further verify the model's robustness and generalization. Besides, the experiments were also conducted on the standard Resume, which contains the standard training set, development set, and testing set, and is widely used in Chinese NER tasks. The experimental results on the extended AgCNER and standard Resume are listed in Table 7. Taking AgCNER as an example, the models that integrated

fine-tuned BERT and adversarial training outperformed the state-of-the-art models, i.e., FGN, Flat-Lattice, and TENER, and delivered significantly better Precision, Recall, and F_1-*scores* on both development and testing sets. The same conclusion could also be drawn on Resume.



(**a**) Ft-BERT-BiLSTM (**b**) ACE-ADP

Figure 6. Confusion matrixes of (**a**) Ft-BERT-BiLSTM, and (**b**) ACE-ADP on AgCNER dataset. *x*-axis: predicted labels; *y*-axis: true-axis; numbers on the cell where x = y represents the TP values.

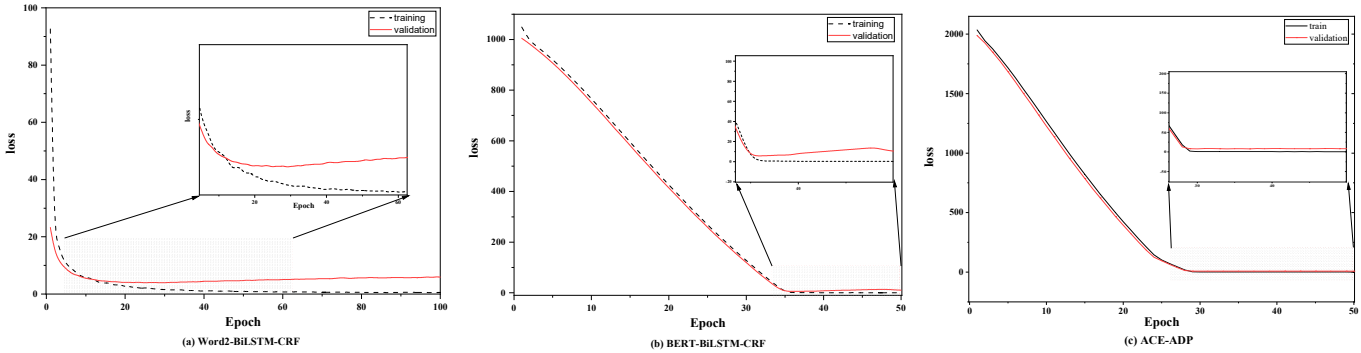


Figure 7. The training and validation losses for (**a**) Word2vec-BiLSTM-CRF, (**b**) BERT-BiLSTM-CRF, and (**c**) ACE-ADP on the AgCNER dataset.

Table 7. Experimental results on extended AgCNER and Resume.

Algorithms	AgCNER						Resume					
	Dev			Test			Dev			Test		
	P	*R*	F_1	*P*	*R*	F_1	*P*	*R*	F_1	*P*	*R*	*F*
ACE-ADP	98.30	98.32	98.31	98.50	98.47	98.49	96.43	97.79	97.11	96.63	97.66	97.14
IDCNN	98.15	98.30	98.23	98.18	98.25	98.21	95.52	96.69	96.10	94.55	96.13	95.33
Gated CNN	98.07	98.76	98.42	98.17	98.75	98.46	96.57	98.47	97.51	96.86	99.00	97.92
RD_CNN	98.71	99.20	98.95	98.69	99.19	98.94	96.87	98.65	97.75	97.11	98.93	98.01
AR-CCNER	97.38	98.03	97.70	97.80	97.88	97.84	95.91	97.79	96.84	97.10	98.33	97.71
CNN-BiLSTM-CRF	97.51	97.81	97.66	97.63	97.58	97.61	95.91	96.38	96.14	95.64	96.79	96.22
FGN	94.33	94.56	94.45	94.26	94.62	94.44	93.13	95.82	94.46	92.12	94.73	93.41
Flat-Lattice	93.52	94.31	93.91	93.71	94.11	93.91	94.74	96.26	95.49	94.90	95.83	95.36
TENER	92.88	95.09	93.97	93.03	95.09	94.05	94.45	95.09	94.77	93.71	94.52	94.11

"dev" represents the development set, "test" donates as the testing set.

Therefore, the above experimental results demonstrated that our works, i.e., integrating the contextual embedding and adversarial training, may contribute to alleviating the over-fitting problem and enhancing the robustness and generalization of the NER models, which would provide us with a feasible solution for the issue of agricultural diseases and pests named entity recognition.

4.3. Convergence

As shown in Figure 8, to evaluate the impact of contextual embeddings and adversarial training on convergence, we took AgCNER as an example and visualized the change curves of F_1 with each iteration by using the BiLSTM that integrated the word2vec, original BERT (Or), fine-tuning BERT (Ft), and adversarial training (AT), respectively. The word2vec-based model showed the slowest convergence speed. Meanwhile, the adversarial training (Word2vec-AT) seems to speed up the convergence of the model a lot at the beginning, but the effect was limited. In contrast, from the change curve of BERT-Or, it could be seen that BERT significantly accelerated the convergence, for the reason that compared to word2vec, the contextual embeddings generated by BERT contain deeper semantic information and would provide better initialization for the model [23]. However, due to the task independence and lack of domain knowledge, the F_1 was slightly lower than that of the word2vec-based model when it tended to be stable. Fine-tuning could equip BERT with domain awareness and provide abundant domain-specific features for the contextual encoders. Therefore, the convergence of BERT-Ft was greatly accelerated; its F_1-*scores* are generally higher than word2vec- and original BERT-based models. In terms of BERT-Or-AT and BERT-Ft-AT, adversarial training could not only accelerate the convergence speed but also significantly improve the recognition performance of the model in the case of high-quality text representation. For example, during the entire training process, the values of F_1 of BERT-Ft-AT were always higher than those of BERT-Ft. Therefore, the above experimental results showed that fine-tuning BERT and adversarial training could improve not only the NER performance but also accelerate the convergence.

4.4. Visualization of Features

To intuitively illustrate the effective effect of BERT on the agricultural and other domain-specific text representation, we visualized the sentence-level embeddings produced by the embeddings-based methods on the training data of the four datasets from From four different perspectives. As shown in Table 8, 100 samples for each dataset were randomly selected, and each sentence-level embedding was projected into a three-dimensional vector by using T-SNE [55]. All the images were obtained by rotating clockwise about the *Z*-axis by 0°, 90°, 180°, and 270°, respectively. In the first row of Table 8, all data points are mixed indiscriminately in space, illustrating that the text representation generated by word2vec cannot effectively represent the semantic features in different domains. In the second row of Table 8, the same type of data points, especially those belonging to AgCNER, CCKS2017, and Resume, were clustered well. However, the data points belonging to CLUENER were relatively loose, and there were several data points mixed with other types of data points, confirming that the original BERT does have a positive effect on the text representation, but due to the task independence, it may not be enough that only using it to generate the domain-specific embeddings. For the fine-tuned BERT (i.e., the last row of Table 8), the data points were correctly divided into four clusters, and the similar data points were more closely distributed, verifying that fine-tuning makes BERT have domain-awareness. Moreover, compared to Word2vec and Original BERT in Table 8, it was more clear of the boundaries between the different types of data points, which was consistent with [9], indicating that injecting domain-specific knowledge by fine-turning may be helpful to the domain-specific NER task.

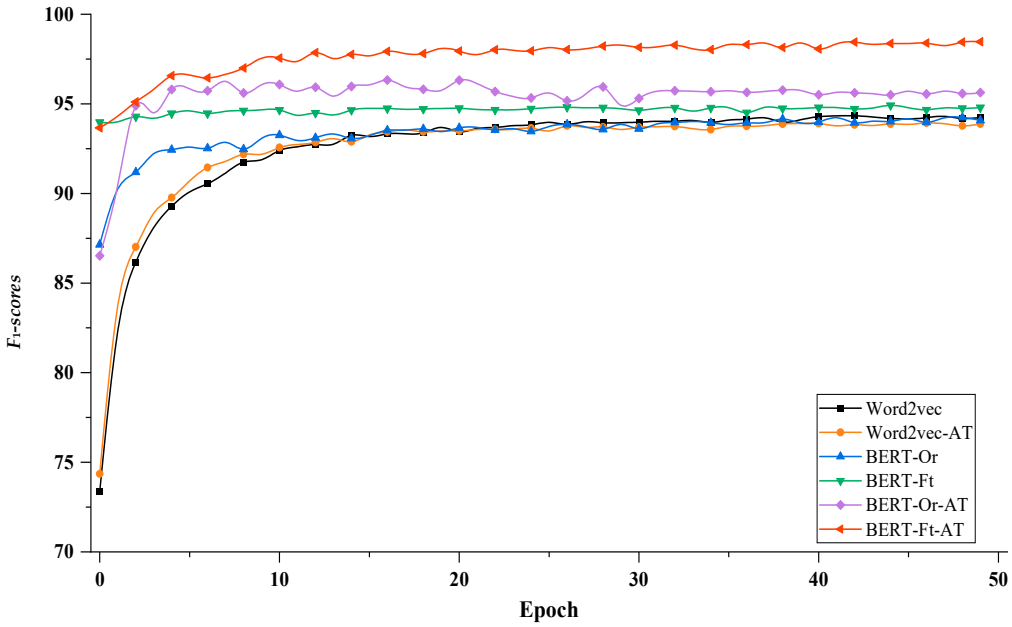
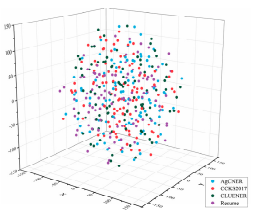
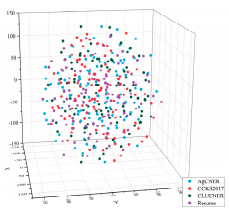
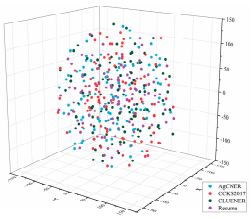
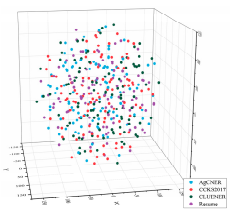
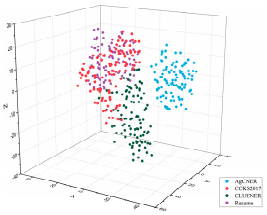
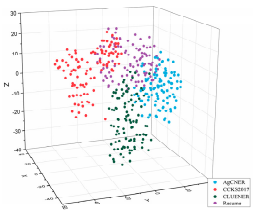
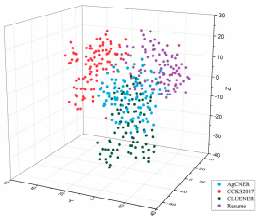
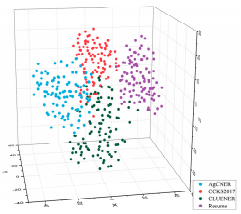
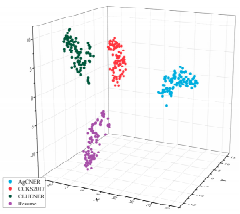
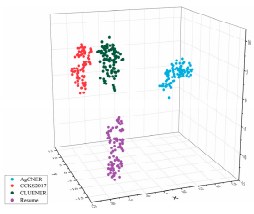
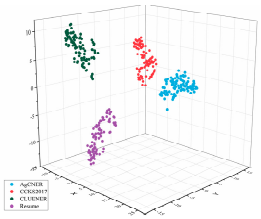
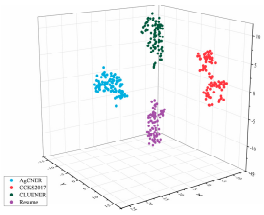


Figure 8. The trend of F_1 for BiLSTM-based models on AgCNER.

Table 8. Visualized results for each dataset.

Rotation Angles	0°	90°	180°	270°
Word2vec				
Original BERT				
Fine-tuned BERT				

4.5. Parameter Analysis

Perturbation size γ represents the degree of perturbation to embedding representation. It is one of the most important parameters during the adversarial training process. According to [56,57], a group of γ range from 0.001 to 0.1 were selected as the candidate perturbation sizes to find the most suitable γ. Similar to [57], the bigger γ was not considered since the larger perturbation may destroy the semantic information of the text representation. As shown in Figure 9, the model achieved different F_1-*scores* for AgCNER and Resume when the perturbation size γ was set to different values, which proved

that different perturbation sizes could affect the performance of the model in different degrees. Besides, the final results were not obvious when γ was varied from 0.001 to 0.01, indicating that when $\gamma < 0.01$, the effect of adversarial training on the NER model was hardly observed. We could also find that when $\gamma = 0.1$, the model obtained the optimal F_1-*scores* of 98.31% and 96.83% on AgCNER and Resume, respectively, indicating that adversarial training could give full play to its performance. Thereby, 0.1 was selected as the perturbation factor during the entire experiment.

In summary, comprehensive experiments and discussions demonstrated the effective performance of ACE-ADP in identifying the agricultural named entities. Besides, an ablation study further demonstrated that it could effectively identify rare entities while accelerating convergence. In the future, we will extend our model to other specific fields to further verify its robustness and generalization. Moreover, we also attempt to improve the model to make it suitable for relation extraction and the joint intent recognition and slot filling task so that to play a role in the construction of agricultural diseases and pests question answering systems based on the knowledge graph.

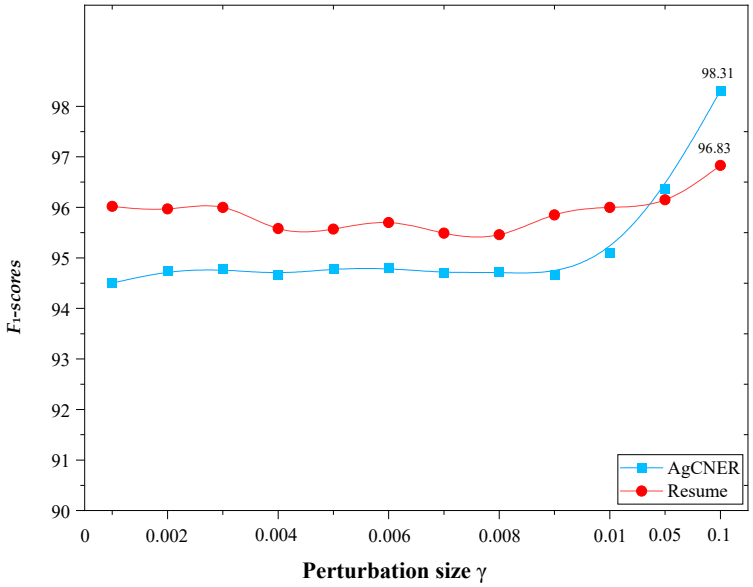


Figure 9. Results by different perturbation sizes γ.

5. Conclusions

To address the problems of rare entity recognition and polysemous words in CNER-ADP tasks, we presented a universal ACE-ADP framework, which effectively enhances the semantic feature representation by introducing and integrating contextual embeddings and adversarial training to recognize the named entities in agricultural diseases and pests. The high-quality context embeddings with agricultural knowledge and high-level features were generated by adopting the BERT that fine-tuned on the agricultural corpus. Furthermore, adversarial training was also introduced to enhance the robustness and generalization of the NER model. Comprehensive experimental results showed that ACE-ADP could significantly improve the F_1-*scores* of the agriculture-related dataset. Moreover, the ablation study and discussion not only verified that ACE-ADP maintained strong robustness and generalization but also showed that it had a strong ability to recognize the rare entities, which is of great benefit to the construction of agricultural diseases and pests question answering systems based on the knowledge graph. To serve digital agriculture, the proposed model will be integrated into the knowledge graph-based question answering systems so that to improve the accuracy in identifying the agricultural diseases and pets-related nouns and make the question answering systems provide more accurate solutions for the agricultural diseases and pests control.

Author Contributions: Conceptualization, methodology, X.G. and X.H.; validation, Z.T., L.D., Z.B. and S.L.; supervision, project administration, L.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program, grant number 2016YFD0300710.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

Appendix A

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

Algorithm A1 Pseudocode for domain-specific named entity recognition task with adversarial training and contextual embeddings.

Input: Fine-tuned BERT model for a specific field, global learning rate l_r, perturbation size γ, the number of iterations T, a domain-specific sentence S, and their ground-truth labels Y.
Output: the predicted labels $\hat{Y}$, the training weights of the model $\hat{\theta}$.

1: Converting the sentence S into the contextual embeddings $E = (e_1, e_2, \ldots, e_n)$ by fine-tuned BERT on the texts in the field of agricultural diseases and pests.
2: For t = 1, . . . , T do
3: $H = BiLSTM(E)$, according to Equation (8) to Equation (10).
4: $P = HW + b$, according to Equation (12).
5: Calculating the $loss(\hat{Y}, Y)$ by using the CRF algorithm.
6: $g \leftarrow \nabla_E log p(\hat{y}|E; \hat{\theta})$,
7: $\varepsilon \leftarrow \gamma\sqrt{d}$
8: $\eta_{adv} \leftarrow -\varepsilon \times g/l2\_normalize(g)$
9: $E_{adv} = E + \eta_{adv}$
10: $loss(\hat{Y}_{adv}, Y) \leftarrow$ Repeat lines 3–9
11: $loss = loss(\hat{Y}, Y) + loss(\hat{Y}_{adv}, Y)$
12: F_1-*scores* $\leftarrow$ conlleval(Y, $\hat{Y}$), calculating the overall F_1-*scores* for predicted labels.
13: If $F_{1\text{-}max} > F_1$-*scores* then
14: $F_{1\text{-}max} \leftarrow F_1$-*scores*
15: Save the weights $\hat{\theta}$ of the model
16: end for
17: Output: the best-predicted labels $\hat{Y}$, the best training weights of the model $\hat{\theta}$.

1. The sentence is converted into contextual embeddings by using BERT, which is fine-tuned on the texts of agricultural diseases and pests.
2. The character-level embeddings are used as input of the BiLSTM to extract the global context features. Note that other contextual encoders such as Gated CNN and RD_CNN can also be used to extract the context features according to the experimental results in Section 3.2.3.
3. The possible labels are predicted, and the loss is calculated by the CRF layer.
4. Calculating the perturbation according to Equation (17) and adding it to the original character-level embeddings.
5. Steps (1) to (4) are repeated until a maximum iteration is reached.

References

1. Lu, J.; Tan, L.; Jiang, H. Review on Convolutional Neural Network (CNN) Applied to Plant Leaf Disease Classification. *Agriculture* **2021**, *11*, 707. [CrossRef]
2. Molina-Villegas, A.; Muñiz-Sanchez, V.; Arreola-Trapala, J.; Alcántara, F. Geographic Named Entity Recognition and Disambiguation in Mexican News using Word Embeddings. *Expert Syst. Appl.* **2021**, *176*, 114855. [CrossRef]
3. Yin, M.; Mou, C.; Xiong, K.; Ren, J. Chinese clinical named entity recognition with radical-level feature and self-attention mechanism. *J. Biomed. Inform.* **2019**, *98*, 103289. [CrossRef]
4. Huang, K.; Altosaar, J.; Ranganath, R. ClinicalBert: Modeling clinical notes and predicting hospital readmission. *arXiv* **2019**, arXiv:1904.05342.
5. Francis, S.; Van Landeghem, J.; Moens, M.F. Transfer learning for named entity recognition in financial and biomedical documents. *Information* **2019**, *10*, 248. [CrossRef]
6. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2019**, *36*, 1234–1240. [CrossRef]
7. 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]
8. Yasunaga, M.; Kasai, J.; Radev, D. Robust multilingual part-of-speech tagging via adversarial training. *arXiv* **2017**, arXiv:1711.04903.
9. Du, C.; Sun, H.; Wang, J.; Qi, Q.; Liao, J. Adversarial and domain-aware bert for cross-domain sentiment analysis. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online. 5–10 July 2020; pp. 4019–4028.
10. Xu, J.; Wen, J.; Sun, X.; Su, Q. A discourse-level named entity recognition and relation extraction dataset for chinese literature text. *arXiv* **2017**, arXiv:1711.07010.
11. Malarkodi, C.S.; Lex, E.; Devi, S.L. Named Entity Recognition for the Agricultural Domain. *Res. Comput. Sci.* **2016**, *117*, 121–132.
12. Nadeau, D.; Sekine, S. A survey of named entity recognition and classification. *Lingvisticae Investig.* **2007**, *30*, 3–26. [CrossRef]
13. Liu, W.; Yu, B.; Zhang, C.; Wang, H.; Pan, K. Chinese Named Entity Recognition Based on Rules and Conditional Random Field. In Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence, ShenZhen, China, 8–10 December 2018; pp. 268–272.
14. WANG Chun-yu, W.F. Study on recognition of chinese agricultural named entity with conditional random fields. *J. Hebei Agric. Univ.* **2014**, *37*, 132–135. [CrossRef]
15. Zhao, P.; Zhao, C.; Wu, H.; Wang, W. Named Entity Recognition of Chinese Agricultural Text Based on Attention Mechanism. *Nongye Jixie Xuebao/Trans. Chin. Soc. Agric. Mach.* **2021**, *52*, 185–192. [CrossRef]
16. Saleem, M.H.; Potgieter, J.; Arif, K.M. Plant Disease Detection and Classification by Deep Learning. *Plants* **2019**, *8*, 468. [CrossRef]
17. Hasan, R.I.; Yusuf, S.M.; Alzubaidi, L. Review of the State of the Art of Deep Learning for Plant Diseases: A Broad Analysis and Discussion. *Plants* **2020**, *9*, 1302. [CrossRef] [PubMed]
18. Zhao, S.; Peng, Y.; Liu, J.; Wu, S. Tomato Leaf Disease Diagnosis Based on Improved Convolution Neural Network by Attention Module. *Agriculture* **2021**, *11*, 651. [CrossRef]
19. Chen, S.; Zhang, K.; Zhao, Y.; Sun, Y.; Ban, W.; Chen, Y.; Zhuang, H.; Zhang, X.; Liu, J.; Yang, T. An Approach for Rice Bacterial Leaf Streak Disease Segmentation and Disease Severity Estimation. *Agriculture* **2021**, *11*, 420. [CrossRef]
20. Hao, X.; Jia, J.; Gao, W.; Guo, X.; Zhang, W.; Zheng, L.; Wang, M. MFC-CNN: An automatic grading scheme for light stress levels of lettuce (Lactuca sativa L.) leaves. *Comput. Electron. Agric.* **2020**, *179*, 105847. [CrossRef]
21. Biswas, P.; Sharan, A. A Noble Approach for Recognition and Classification of Agricultural Named Entities using Word2Vec. *Int. J. Adv. Stud. Comput. Sci. Eng.* **2021**, *9*, 1–8.
22. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
23. Jawahar, G.; Sagot, B.; Seddah, D. What Does BERT Learn about the Structure of Language? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 6 July 2019; pp. 3651–3657.
24. Zhang, S.; Zhao, M. Chinese agricultural diseases named entity recognition based on BERT-CRF. In Proceedings of the 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), Harbin, China, 25–27 December 2020; pp. 1148–1151.
25. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF models for sequence tagging. *arXiv* **2015**, arXiv:1508.01991.
26. Strubell, E.; Verga, P.; Belanger, D.; McCallum, A. Fast and Accurate Entity Recognition with Iterated Dilated Convolutions. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 1 September 2017; pp. 2670–2680.
27. Dauphin, Y.N.; Fan, A.; Auli, M.; Grangier, D. Language modeling with gated convolutional networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 933–941.
28. Qiu, J.; Wang, Q.; Zhou, Y.; Ruan, T.; Gao, J. Fast and accurate recognition of Chinese clinical named entities with residual dilated convolutions. In Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain, 3–6 December 2018; pp. 935–942.
29. Yan, H.; Deng, B.; Li, X.; Qiu, X. Tener: Adapting transformer encoder for named entity recognition. *arXiv* **2019**, arXiv:1911.04474.

30. Chen, H.; Lin, Z.; Ding, G.; Lou, J.; Zhang, Y.; Karlsson, B. GRN: Gated relation network to enhance convolutional neural network for named entity recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 6236–6243.
31. Zhu, Y.; Wang, G. CAN-NER: Convolutional Attention Network for Chinese Named Entity Recognition. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 3384–3393.
32. Li, L.; Zhao, J.; Hou, L.; Zhai, Y.; Shi, J.; Cui, F. An attention-based deep learning model for clinical named entity recognition of Chinese electronic medical records. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 1–11. [CrossRef]
33. Cao, P.; Chen, Y.; Liu, K.; Zhao, J.; Liu, S. Adversarial transfer learning for Chinese named entity recognition with self-attention mechanism. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 2–4 November 2018; pp. 182–192.
34. Wang, C.; Chen, W.; Xu, B. Named entity recognition with gated convolutional neural networks. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 110–121.
35. Li, X.; Yan, H.; Qiu, X.; Huang, X.-J. FLAT: Chinese NER Using Flat-Lattice Transformer. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Seattle, WA, USA, 5–10 July 2020; pp. 6836–6842.
36. Cetoli, A.; Bragaglia, S.; O'Harney, A.; Sloan, M. Graph Convolutional Networks for Named Entity Recognition. In Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories, Prague, Czech Republic, 1 July 2017; pp. 37–45.
37. Gui, T.; Zou, Y.; Zhang, Q.; Peng, M.; Fu, J.; Wei, Z.; Huang, X.-J. A lexicon-based graph neural network for chinese ner. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3 November 2019; pp. 1039–1049.
38. Li, J.; Sun, A.; Han, J.; Li, C. A survey on deep learning for named entity recognition. *IEEE Trans. Knowl. Data Eng.* **2020**. [CrossRef]
39. Pre-trained models for natural language processing: A survey. *arXiv* **2020**, arXiv:2003.08271.
40. Zhang, R.; Lu, W.; Wang, S.; Peng, X.; Yu, R.; Gao, Y. Chinese clinical named entity recognition based on stacked neural network. *Concurr. Comput. Pract. Exp.* **2020**, e5775. [CrossRef]
41. Suman, C.; Reddy, S.M.; Saha, S.; Bhattacharyya, P. Why pay more? A simple and efficient named entity recognition system for tweets. *Expert Syst. Appl.* **2021**, *167*, 114101. [CrossRef]
42. Yang, Z.; Chen, H.; Zhang, J.; Ma, J.; Chang, Y. Attention-based multi-level feature fusion for named entity recognition. *IJCAI Int. Jt. Conf. Artif. Intell.* **2020**, *2021*, 3594–3600. [CrossRef]
43. Liu, X.; Zhou, Y.; Wang, Z. Deep neural network-based recognition of entities in Chinese online medical inquiry texts. *Futur. Gener. Comput. Syst.* **2021**, *114*, 581–604. [CrossRef]
44. Chiu, J.P.C.; Nichols, E. Named entity recognition with bidirectional LSTM-CNNs. *Trans. Assoc. Comput. Linguist.* **2016**, *4*, 357–370. [CrossRef]
45. Ma, X.; Hovy, E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 3 August 2016; pp. 1064–1074.
46. Peters, M.E.; Ruder, S.; Smith, N.A. To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks. In Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), Florence, Italy, 5 August 2019; pp. 7–14.
47. Song, C.H.; Sehanobish, A. Using Chinese Glyphs for Named Entity Recognition (Student Abstract). In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 13921–13922.
48. Xuan, Z.; Bao, R.; Jiang, S. FGN: Fusion glyph network for Chinese named entity recognition. *arXiv* **2020**, arXiv:2001.05272.
49. Xu, L.; Dong, Q.; Liao, Y.; Yu, C.; Tian, Y.; Liu, W.; Li, L.; Liu, C.; Zhang, X. CLUENER2020: Fine-grained named entity recognition dataset and benchmark for chinese. *arXiv* **2020**, arXiv:2001.04351.
50. Zhang, Y.; Yang, J. Chinese NER Using Lattice LSTM. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 1554–1564.
51. Li, S.; Zhao, Z.; Hu, R.; Li, W.; Liu, T.; Du, X. Analogical Reasoning on Chinese Morphological and Semantic Relations. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Melbourne, Australia, 15–20 July 2018; pp. 138–143.
52. Prechelt, L. Early stopping-but when? In *Neural Networks: Tricks of the Trade*; Springer: Berlin/Heidelberg, Germany, 1998; pp. 55–69.
53. Bekoulis, G.; Deleu, J.; Demeester, T.; Develder, C. Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Syst. Appl.* **2018**, *114*, 34–45. [CrossRef]
54. Miyato, T.; Dai, A.M.; Goodfellow, I. Adversarial training methods for semi-supervised text classification. *arXiv* **2016**, arXiv:1605.07725.
55. der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
56. Zhao, S.; Cai, Z.; Chen, H.; Wang, Y.; Liu, F.; Liu, A. Adversarial training based lattice LSTM for Chinese clinical named entity recognition. *J. Biomed. Inform.* **2019**, *99*, 103290. [CrossRef] [PubMed]
57. Liu, X.; Cheng, H.; He, P.; Chen, W.; Wang, Y.; Poon, H.; Gao, J. Adversarial training for large neural language models. *arXiv* **2020**, arXiv:2004.08994.