Agricultural Entity Recognition Based on Semantic Fusion and Model Distillation: Postprint

Authors: Li Liangde, Wang Xiujuan, Kang Mengzhen, Huajing, Menghan Fan

Date: 2023-02-17T00:00:00+00:00

Abstract

Currently, annotated data for agricultural entity recognition is scarce, and some publicly available models rely on handcrafted features, resulting in low accuracy. Although certain agricultural entity recognition models based on deep learning methods have achieved improved performance, they suffer from high inference latency and large parameter counts. This study proposes an agricultural entity recognition method based on knowledge distillation. First, we construct an agricultural knowledge graph using massive agricultural data from the Internet, and obtain weakly annotated corpora through distant supervision. Second, targeting the characteristics of entity recognition, we propose an Attention-based BERT Layer Aggregation model (BERT-ALA) that fuses semantic features from different layers; combined with Bidirectional Long Short-Term Memory networks (BiLSTM) and Conditional Random Fields (CRF), we obtain the BERT-ALA+BiLSTM+CRF model as the teacher model. Finally, we employ the BiLSTM+CRF model as the student model to distill the teacher model, ensuring that prediction time and parameter count meet online service requirements. Experiments on the agricultural entity recognition dataset constructed in this study and two public datasets demonstrate that the BERT-ALA+BiLSTM+CRF model achieves an average macro-F1 improvement of 1% over the baseline BERT+BiLSTM+CRF model. The distilled student model BiLSTM+CRF achieves an average macro-F1 improvement of 3.3% over the model trained on original data, with prediction time reduced by 33% and storage space reduced by 98%. Experimental results validate the effectiveness of the attention mechanism-based BERT layer fusion model and knowledge distillation for agricultural entity recognition.

Full Text

Agricultural Named Entity Recognition Based on Semantic Aggregation and Model Distillation

LI Liangde[12], WANG Xiujuan[13], KANG Mengzhen[12]*, HUA Jing[14], FAN Menghan[12]

[1]The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China
[3]Beijing Engineering Research Center of Intelligent Systems and Technology, Beijing 100190, China
[4]Qingdao Smart AgriTech Co., Ltd., Qingdao 266000, China

Abstract: With the development of smart agriculture, automatic question answering (Q&A) systems for agricultural knowledge are needed to improve the efficiency of agricultural information acquisition. Agricultural named entity recognition plays a key role in such systems, as it helps obtain information, understand agricultural questions, and provide answers from knowledge graphs. However, due to the scarcity of labeled agricultural named entity data, some existing open agricultural entity recognition models rely on manual features, resulting in low accuracy. While certain agricultural entity recognition models based on deep learning have improved performance, they suffer from high inference latency and large parameter counts. This study proposes an agricultural entity recognition method based on knowledge distillation. First, massive agricultural data from the Internet were leveraged to construct an agricultural knowledge graph (AgriKG), which was then used to generate weakly labeled corpora through distant supervision. Second, considering the characteristics of entity recognition, an Attention-based Layer Aggregation model (ALA) was proposed to fuse semantic features from different layers; combined with a Bidirectional Long Short-Term Memory network (BiLSTM) and Conditional Random Field (CRF), this formed the BERT-ALA+BiLSTM+CRF teacher model to ensure prediction time and parameter count meet online service requirements. Finally, the BiLSTM+CRF model served as the student model. Experiments conducted on the agricultural entity recognition dataset constructed in this study and two public datasets showed that the macro-F1 of the BERT-ALA+BiLSTM+CRF model improved by 1% compared to the baseline BERT+BiLSTM+CRF model. The distilled student model achieved an average macro-F1 improvement of 3.3% over models trained on original data, while prediction time decreased by 33% and storage space reduced by 98%. These results verify the effectiveness of the attention-based layer aggregation mechanism and knowledge distillation for agricultural entity recognition.

Keywords: distant supervision; agricultural knowledge graph; agricultural Q&A system; named entity recognition; knowledge distillation; deep learning; BERT; BiLSTM

1 Introduction

With the development of agricultural Internet and the generational turnover of agricultural practitioners, rapid dissemination and application of agricultural knowledge are required to address the shortage of agricultural technical experts. Currently, agricultural knowledge Q&A on the Internet is primarily handled by human experts, which is not only inefficient but also limited by the scarcity of expert resources. If computers could understand user-input agricultural questions and provide intelligent answers through agricultural knowledge graphs, the efficiency of agricultural knowledge Q&A would be greatly improved.

Agricultural intelligent Q&A systems include four components: information extraction [1], knowledge graph construction, question understanding, and knowledge base-based Q&A. Information extraction is crucial for understanding questions and answering them based on agricultural knowledge graphs. Named entity recognition [2] identifies entity mentions and their categories in text, serving as a fundamental task in natural language processing. Based on agricultural entity recognition, key information can be extracted from texts to build agricultural knowledge graphs, enabling agricultural knowledge structuring and subsequent Q&A. The Internet stores vast amounts of unstructured agricultural texts, and transforming these disorganized texts into structured agricultural knowledge to build agricultural knowledge graphs is an essential step in implementing agricultural intelligent Q&A systems.

Agriculticultural knowledge data, particularly annotated data, is difficult to obtain, resulting in relatively few studies on agricultural knowledge graph construction and information extraction. Existing agricultural entity recognition solutions often require large amounts of training data, necessitating costly manual annotation of entity recognition data when applied. These models also suffer from issues such as reliance on manual feature extraction and suboptimal entity recognition performance, or they fail to consider practical online requirements for prediction time and model size, remaining at the experimental validation stage. Li and Zhang [3] used dictionaries for entity recognition, building a web knowledge extraction model based on agricultural ontology. However, since web knowledge bases cannot cover all agricultural entities, this approach suffers from low recall. Wang and Wang [4] employed Conditional Random Fields [5] for named entity recognition, but this method requires manual feature construction and has low model capacity, making it difficult to handle complex entity recognition tasks. Malarkodi et al. [6] applied CRF models with syntactic and lexical features, similarly relying on manual feature construction. Liu [7] used Dense Connected Bi-directional Long Short-Term Memory (DC-LSTM) + Conditional Random Field (CRF) architecture for agricultural domain named entity recognition. Due to its multi-layer densely connected structure, this approach has long inference times and many parameters, making it difficult for practical online use. Biswas et al. [8] utilized WordNet [9] for agricultural entity

recognition, which is essentially similar to dictionary matching but expands the dictionary using WordNet' s word correlations.

Currently, both traditional methods based on Conditional Random Fields and deep learning-based [10] entity recognition models are data-driven and require massive annotated data as support. In the agricultural domain, where large amounts of readily available annotated data are lacking, directly applying general domain entity recognition solutions is ineffective. Therefore, this study proposes an agricultural domain data annotation scheme based on distant supervision [11] to address the scarcity of agricultural entity recognition annotated data. Distant supervision was first proposed by Mintz at the 47th Annual Meeting of the Association for Computational Linguistics (ACL), which automatically constructs large-scale training data by aligning knowledge bases with texts, reducing reliance on manually annotated data and enhancing cross-domain adaptation capabilities. It has been widely applied in relation extraction [12]. The motivation behind distant supervision is to solve the problem of difficult-to-obtain relation extraction annotation data, which is analogous to the problem of scarce annotation data in agricultural entity recognition. Therefore, this paper adapts the distant supervision concept to the entity recognition domain. While polysemy poses significant noise challenges for distant supervision in general domains (e.g., mapping "Apple" in "Apple phone" to the fruit "apple"), in specialized domains like agriculture, although missing annotations exist, word semantics are relatively fixed, resulting in less overall noise. Thus, distant supervision is a feasible solution that effectively circumvents the lack of annotated data in agriculture.

This study adopts the popular large-scale pre-trained model in natural language processing, Bidirectional Encoder Representations from Transformers (BERT) [13]. On one hand, pre-trained models trained on massive Internet data have large capacity and can fit complex entity recognition tasks. On the other hand, agricultural entity recognition annotated data is relatively scarce, while pre-trained models trained on large-scale corpora contain substantial basic linguistic knowledge. Fine-tuning on top of large-scale pre-trained models allows agricultural entity recognition models to incorporate this fundamental linguistic knowledge. Additionally, considering the characteristics of agricultural entity recognition, this study proposes an attention-based layer aggregation mechanism for BERT.

Online Q&A systems require models with low time and space complexity. While the BERT-based model proposed earlier performs well, its large parameter count leads to high inference latency, making it difficult to meet real-time inference requirements. Model distillation [14] transfers the "knowledge" generalization ability of a trained complex model to a simpler network, or enables a simple network to learn the "knowledge" from a complex model. The trained complex model is called the teacher model, while the learning simple model is called the student model. Considering online requirements for prediction time and model size, this study uses BiLSTM + CRF [15] as the student model to distill the

previously obtained BERT-based series models.

2 Methods

2.1 Overall Architecture

The agricultural entity recognition architecture proposed in this study mainly includes three modules: a weakly labeled corpus construction module, a model training module, and an online inference module (Figure 1 [Figure 1: see original paper]).

The weakly labeled corpus construction module adopts the distant supervision concept and consists of two stages: first, the agricultural knowledge graph construction stage, where Internet agricultural resources are crawled, filtered to obtain agricultural entities, and used to build an agricultural knowledge graph; second, the data weak labeling stage, where entities from the agricultural knowledge graph are stored in a prefix tree [17] as a dictionary, and forward maximum matching is applied to sentences in the text to obtain weakly labeled entity results. The model training module also contains two stages: first, the teacher model training stage, where the proposed teacher model is trained using weakly labeled data; second, the model distillation stage [14], where a lightweight model serves as the student model to distill the teacher model. The online inference module accepts text sent from the client, merges the results from the dictionary and student model, and returns them to the client.

2.2 Data Sources

Currently, the agricultural domain lacks open-source Chinese agricultural knowledge graphs and agricultural entity recognition corpora. Hudong Baike and Baidu Baike are open-source Chinese encyclopedia websites containing substantial agricultural entities and knowledge. Much agricultural knowledge on various agricultural websites is also similar to that on encyclopedia websites, and different encyclopedia sites contain similar agricultural knowledge. Considering that Hudong Baike is easier to crawl than other encyclopedia sites and open-source agricultural information websites, this study chose to crawl Hudong Baike data to establish an agricultural knowledge graph for building the knowledge graph and annotating entity recognition training corpora. Documents corresponding to agricultural entities in the Hudong Baike database were segmented into sentences to obtain agricultural entity recognition corpora.

2.3 Weakly Labeled Agricultural Named Entity Recognition Corpus via Distant Supervision

Applying the distant supervision concept to entity recognition assumes that if a word in a sentence has the same name or alias as an entity in the knowledge graph, then that word corresponds to the entity in the knowledge graph. The distant supervision concept has two issues: first, polysemous entities can lead to

annotation errors (e.g., mapping “Apple” in “Apple phone” to the fruit “apple”), but polysemy can be ignored in specialized domain texts like agriculture; second, entities not present in the agricultural knowledge graph suffer from missing annotations. The weak labeling of texts through distant supervision can be divided into two stages: first, crawling Internet-based collaborative writing systems (Wiki) to build an agricultural knowledge graph, and applying rule matching to Wiki ontology tag information to infer entity types, filtering entities with types such as crops, diseases, and pesticides; second, weakly labeling the corpora by storing agricultural knowledge graph entities in a prefix tree as a dictionary and performing forward maximum matching on sentences in the text to obtain weakly labeled entity results.

For example, the sentence “How to transplant tomato seedlings” can be forward maximum matched to obtain the word “tomato” corresponding to the “tomato” entity in the agricultural knowledge graph, where the tomato entity category is crop. This generates the label sequence: O (“how”) O (“to”) O (“transplant”) O (“tomato”) B_{crop} (“to”) I_{crop} (“mato”) O (“seed”) O (“lings”). Here, O (other) represents non-entities, B (begin) represents entity start positions, I (interior) represents entity interior and end positions, and crop indicates the entity type is crop. B_{crop} I_{crop} represents an entity of type crop, corresponding to the start and end positions of the entity, which are the 4th and 5th words (“tomato”) in the sentence.

2.4 Teacher Model

Deep learning models + Conditional Random Fields [15,18,19] are mainstream models for named entity recognition [15]. Deep learning models refer to models like BiLSTM [20] and BERT [13] that extract semantic features from texts, obtaining probabilities from words to each entity category; Conditional Random Fields calculate transition probabilities between entity categories, combining generation probabilities and transition probabilities for end-to-end training.

2.4.1 BERT Model The BERT model, released by Google AI in 2018, achieved state-of-the-art results on 11 different natural language processing validation tasks. Simply put, BERT trained a general language understanding model on massive text corpora using self-supervised methods, then set lightweight downstream task interfaces on this model to perform specific natural language processing tasks. The BERT model structure is shown in Figure 2 [Figure 2: see original paper].

The BERT model mainly consists of three parts: the input layer, multiple transformer encoders, and the output layer. The input layer comprises token embedding, position embedding, and segment embedding. Token embedding segments text into words and converts words into vectors; position embedding encodes word position information into feature vectors, enabling the model to obtain word position information; segment embedding distinguishes between two input sentences. The transformer encoder [21] uses self-attention mechanisms to

enable word-to-word interaction and obtain sentence semantic representations. The output layer determines specific structures based on downstream tasks on top of sentence semantic representations. BERT training consists of two stages: pre-training and fine-tuning. The pre-training stage uses self-supervised training, with the main task being Masked Language Model, which randomly masks certain words in sentences and predicts them. This process requires no annotated corpora and can directly obtain data from the Internet. In the fine-tuning stage, different output layers and objective functions are set for specific tasks, and model parameters are further updated using small amounts of annotated data to complete domain-specific model training.

2.4.2 Long Short-Term Memory Network Long Short-Term Memory (LSTM) networks [20] use gate mechanisms to improve the gradient vanishing problem in Recurrent Neural Networks (RNN). Bidirectional LSTM (BiLSTM) consists of two unidirectional LSTM networks, one propagating forward through time and the other backward. For text sequences, BiLSTM can effectively capture contextual information and is effective for sequence labeling tasks like entity recognition.

2.4.3 Conditional Random Field Conditional Random Field (CRF) models [5] are probabilistic undirected graph models that can solve sequence labeling tasks. Given an observation sequence X, the probability of the hidden state sequence Y is P(Y|X). The CRF used in named entity recognition is mainly the linear-chain CRF, with the mathematical formula shown below:

$$P(y|x) = \frac{\exp\left(\sum_{k=1}^{K} w_k f_k(y, x)\right)}{Z(x)}$$

where f_k is the feature function, w_k is the weight of the feature function, and $Z(x)$ is the normalization factor. During prediction, the model uses the Viterbi algorithm, a dynamic programming algorithm that finds the label sequence Y with maximum probability given observation sequence X and parameters.

2.4.4 Deep Learning Model + Conditional Random Field Deep learning models essentially treat deep models as text feature extractors to obtain text features, which are then passed through a fully connected layer to obtain scores from words to entity categories, denoted as P, and input into the CRF layer. The CRF layer contains a transition matrix A, representing the transition scores between two tags. The model scores a sentence x with labels equal to y, and the score is softmax-normalized to obtain probability, expressed as follows:

$$\text{score}(x, y) = \sum_{i=1}^{n+1} P_{i,y_i} + A_{y_{i-1},y_i}$$

$$P(y|x) = \frac{\exp(\text{score}(x, y))}{\sum_{y'} \exp(\text{score}(x, y'))}$$

The entire sentence' s score equals the sum of scores at each position, where each position' s score is determined by the deep learning model' s output P and transition score A. During training, the model maximizes the log loss function.

Deep models can be BERT, BiLSTM, Iterated Dilated Convolutional Neural Network (IDCNN), etc. Currently, BERT and BiLSTM are most commonly used in entity recognition. BiLSTM+CRF was proposed by Dong et al. [22] in 2016 for general domain named entity recognition; BERT+CRF was proposed by Souza et al. [19] for Portuguese named entity recognition. However, BERT' s transformer self-attention mechanism can destroy relative position information [23]. To address BERT' s insufficient relative position information extraction capability, one approach uses BERT+BiLSTM [22] as the deep model. BERT+BiLSTM+CRF was proposed by Jiang et al. [24] for general domain named entity recognition. BERT serves to provide dynamic word vectors, while BiLSTM models relative position information. Therefore, this study sets up three baseline models: BiLSTM+CRF [22], BERT+CRF [19], and BERT+BiLSTM+CRF [24] for agricultural entity recognition experiments, selecting the model with better experimental results as the teacher model to distill a lightweight student model. These three baseline models have been validated as effective in other domains.

2.4.5 Attention-Based BERT Layer Aggregation Mechanism Entity recognition tasks have high requirements for low-level syntactic and semantic features but relatively lower requirements for high-level semantic features. BERT is a multi-layer transformer [21] feature extractor, with the BERT-base model containing 12 layers. Multi-layer transformers slow down model inference speed. On the other hand, Jawahar et al. [25] pointed out in their ACL 2019 paper that BERT' s lower layers learn phrase-level information representation, middle layers learn rich linguistic features, and higher layers learn rich semantic information features. For general domain entity recognition, models focus on high-level semantic features while neglecting the low-level features urgently needed for entity recognition tasks. For vertical domains like agriculture, determining entity boundaries is more difficult than determining entity categories because entity meanings in vertical domains are relatively easier to discriminate than in general domains. Therefore, phrase-level information representation contained in low-level features is more important for discriminating entity boundaries, and considering only high-level semantic information is clearly unreasonable. Additionally, the amount of annotated data obtained through distant supervision in this study is limited, and directly using high-level information can easily lead to overfitting. Therefore, this study proposes an attention-based BERT layer aggregation mechanism. BERT models contain multiple transformer encoder layers, with different BERT model sizes having different numbers of layers, generally

12, 24, or 48. Denoting BERT' s number of layers as L, attention-based layer aggregation is performed, where α and γ are trainable parameters, as shown in formulas (5) and (6):

$$h = \gamma \sum_{i=1}^{L} w_i h_i$$

$$w_i = \frac{\exp(\alpha_i)}{\sum_{j=1}^{L} \exp(\alpha_j)}$$

where h is the output of BERT model intermediate layers, and w is the weight of each layer.

This study names the attention-based BERT layer aggregation model BERT-ALA (Attention-Based Layer Aggregation for BERT), and uses this name uniformly in subsequent experiments. This mechanism can be applied to any BERT-based model. Applying BERT-ALA to BERT+BiLSTM+CRF yields BERT-ALA+BiLSTM+CRF, with the main structure shown in Figure 3 [Figure 3: see original paper]. BERT model outputs from different layers are weighted by a set of learnable weight parameters to obtain the final feature representation, which is then fed into subsequent BiLSTM and CRF layers for entity recognition.

2.5 Model Distillation

Model distillation [14] transfers the generalization ability "knowledge" of a trained complex model to a simpler network, or enables a simple network to learn the "knowledge" from a complex model. The trained complex model is called the teacher model, while the learning simple model is called the student model. While several BERT-based models were proposed earlier, BERT' s large parameter count leads to high inference latency, making it difficult to meet real-time inference demands. Therefore, this study uses BiLSTM+CRF as the student model to distill the previously proposed teacher model. Unlike traditional model distillation that only distills the final layer output, this study also distills the intermediate BiLSTM layer of the teacher model. The distillation loss function consists of three terms, with the objective function expressed as:

$$\text{loss} = \alpha_1 \text{MSE}_{\text{loss}}(h^{(T)}_{\text{BiLSTM}}, h^{(S)}_{\text{BiLSTM}}) + \alpha_2 \text{CE}_{\text{loss}}(h^{(T)}_{\text{CRF}}, h^{(S)}_{\text{CRF}}) + \alpha_3 \text{CRF}_{\text{loss}}(y_{\text{true}}, h^{(S)}_{\text{CRF}})$$

where S denotes the student model, T denotes the teacher model, and h_{layer} represents the layer output (BiLSTM layer, CRF layer). The three distillation loss terms respectively represent: (1) the student model' s BiLSTM layer output fitting the teacher model' s BiLSTM layer output via MSE; (2) the cross-entropy between the probability distributions output by the student model' s CRF layer

and the teacher model' s CRF layer; (3) the original CRF loss [15], which is calculated from the CRF layer output probability and the true entity recognition labels.

2.6 Model Inference

During the inference stage, after receiving text input from the client, the process includes three stages: (1) Dictionary matching identifies agricultural-type entities S_1 in the sentence. (2) The student model predicts agricultural entities S_2 in the sentence. (3) The annotation results from the model and dictionary are aggregated using union and returned to the client. For entities present in S_2 but not in S_1, which do not exist in the dictionary, they are returned for manual expert review. Confirmed new words are added to the dictionary to improve dictionary coverage.

3 Experimental Validation and Analysis

3.1 Evaluation Metrics

The experiment adopts exact matching mode. Mentions recognized by the entity recognition model and entities in the ground truth are both represented as $(\text{start}, \text{end}, \text{type})$, where start and end represent mention or entity boundaries, and type represents the category. For the entity recognition domain, TP, FP, and FN are defined as follows:

1. **True Positive (TP):** Mentions recognized by the agricultural entity recognition model that correspond to entities in the ground truth.
2. **False Positive (FP):** Mentions recognized by the agricultural entity recognition model that do not correspond to entities in the ground truth, including cases where boundaries are correct but types are incorrect.
3. **False Negative (FN):** Entities in the ground truth that are not recognized by the agricultural entity recognition model.

Based on the above definitions, Precision, Recall, and F1-score are calculated as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Entities contain multiple types, and F1 is calculated separately for different entity types. The overall F1 uses macro-F1 [26], calculated as the average of F1 scores across categories:

$$\text{macro-F1} = \frac{1}{N}\sum_{i=1}^{N}\text{F1}_i$$

3.2 Experimental Design

This study selected two domains—agriculture and medicine—with three datasets for experimental validation. The medical domain was chosen because, like agriculture, it is a specialized domain with relatively more research and readily available open-source entity recognition annotated data.

Dataset 1: Agricultural domain text obtained from Hudong Baike, segmented by sentences. The training set was constructed using distant supervision, while the validation set was manually annotated. It contains 4,662 crop entities and 695 disease entities. The training and test sets are split 8:2, with 10,277 training samples and 2,532 test samples. The dataset has been open-sourced on the data modeling and analysis competition platform Kaggle (https://www.kaggle.com/supportvectordevin/agriculture-pedia).

Dataset 2: From the "Agricultural Q&A Data Processing Challenge" on the iFLYTEK open platform (http://challenge.xfyun.cn/topic/info?type=agriculture) [25], which annotates named entity labels for crops, pests, and pesticides. The dataset contains 100,660 pest entities, 250,740 pesticide entities, and 5,796 crop entities. The training set includes 15,624 samples, and the test set includes 3,906 samples.

Dataset 3: Medical domain data from CCKS 2017 task 2, Clinical Named Entity Recognition (CNER) (https://github.com/zjy-ucas/ChineseNER). For a given set of electronic medical record documents (plain text files), the task aims to identify and extract clinically relevant entity names. The dataset contains 12,821 symptom and sign entities, 17,655 examination and test entities, 4,560 disease and diagnosis entities, 4,940 treatment entities, and 17,556 body part entities. The training set includes 10,787 samples, and the test set includes 2,697 samples.

For model hyperparameters, LSTM+CRF uses fastText Chinese word embeddings [27] with 128 LSTM hidden layers. Training uses the Adam optimizer [28], with a BERT layer learning rate of 10^{-5} and other layers at 10^{-3}. Batch size is 32, with each batch padded to the longest sentence within the batch to reduce memory consumption, but with a maximum truncation length of 64.

3.3 Baseline Model Comparison

The macro-F1 of three baseline models was tested on the three datasets, with results shown in Table 1 .

Table 1 Comparison of macro-F1 with three baseline models

Model	Dataset 1	Dataset 2	Dataset 3
BiLSTM+CRF	-	-	-
BERT+CRF	-	-	-
BERT+BiLSTM+CRF	-	-	-

Results on Dataset 1 show that models trained on distant supervision training data perform well on manually annotated test sets, proving the effectiveness of dataset construction via distant supervision. Introducing the large-scale pre-trained model BERT significantly improves model performance compared to BiLSTM, with macro-F1 improvements of 7.75% on Dataset 1, 13.39% on Dataset 2, and a relatively smaller improvement of 1.64% on Dataset 3 (medical entity recognition data). Adding BiLSTM after BERT can alleviate BERT' s insufficient relative position capture capability to some extent, with macro-F1 improvements of 0.71% on Dataset 1, 0.36% on Dataset 2, and 0.69% on Dataset 3 compared to BERT+CRF.

3.4 Validation of Attention-Based BERT Layer Aggregation Mechanism

For the two BERT-based models (BERT+CRF and BERT+BiLSTM+CRF), the layer aggregation mechanism was applied to improve BERT, and results were validated to determine whether entity recognition performance improved. Results are shown in Table 2 .

Table 2 Validation of layer aggregation mechanism

Model	Dataset 1	Dataset 2	Dataset 3
BERT+CRF	-	-	-
BERT-ALA+CRF	-	-	-
BERT+BiLSTM+CRF	-	-	-
BERT-ALA+BiLSTM+CRF	-	-	-

Validation results demonstrate that the attention-based layer aggregation mechanism improves entity recognition performance across all three datasets, indicating that the layer aggregation mechanism has certain universality in the entity recognition domain. BERT-ALA+CRF and BERT-ALA+BiLSTM+CRF achieve approximately 1% macro-F1 improvement over baseline models. BERT-ALA+BiLSTM+CRF achieves the best performance among all models and is therefore selected as the teacher model to guide the student model learning in the distillation portion. This study primarily applies BERT-ALA+BiLSTM+CRF to the agricultural entity recognition domain.

3.5 Validation of Model Distillation Effect

Through model distillation, the teacher model obtained is BERT-ALA+BiLSTM+CRF, and the student model is BiLSTM+CRF. Compared with the teacher model, the student model shows improvements in time and space complexity. This study uses the average time to predict 1,000 samples to represent model prediction time for comparing student model time complexity improvements; model size is represented by storage space occupied to verify student model space complexity improvements. Since these two metrics are data-independent, experiments were conducted on three datasets and averaged. Results show that the distilled student model reduces prediction time by 33% and model size by 98% compared to the teacher model, demonstrating significant improvements in time and space complexity that make the model more suitable for online prediction scenarios.

This study tested the improvement of the distilled student model over an equivalent model trained with annotated data, with macro-F1 comparison results shown in Table 3 .

Table 3 Comparison of macro-F1 with teacher model and student model

Model	Dataset 1	Dataset 2	Dataset 3
BiLSTM+CRF (Teacher)	-	-	-
BiLSTM+CRF (Student)	-	-	-

Validation results demonstrate that using model distillation as a training method enables the student model to learn more dark knowledge compared to training on the original data. The distilled student model achieves macro-F1 improvements of 3.1% on Dataset 1, 4.09% on Dataset 2, and 2.82% on Dataset 3.

3.6 Student Model Effect Demonstration

The primary application scenario of this study is agricultural entity recognition. Therefore, taking tomato as an example, several Q&A pairs [27] about tomato were selected to validate the final online distilled student model effect. Sentences and their recognition results are as follows:

Question 1: What are the symptoms and control methods for tomato virus disease?
Recognition result: `{'mention': 'tomato virus disease', 'type': 'disease', 'offset': 0}`

Question 2: How does tomato vascular wilt occur and how to prevent it?
Recognition result: `{'mention': 'tomato vascular wilt', 'type': 'disease', 'offset': 0}`

Question 3: Symptoms: Tomato bacterial spot disease mainly harms leaves, stems, flowers, petioles, and fruits.
Recognition result: `{'mention': 'tomato bacterial spot disease', 'type': 'disease', 'offset': 3}`

The entities in Questions 1, 2, and 3 are all completely recognized. Among them, the entities "tomato vascular wilt" and "tomato bacterial spot disease" in Questions 2 and 3 do not appear in the dictionary and are not present in the annotated data, but the model can successfully recognize them, verifying the model' s good generalization performance.

4 Conclusion

This study proposes using distant supervision to construct agricultural entity recognition data, which has the problem of missing annotations. Based on the assumption that sentences with missing annotations are far fewer than correctly annotated sentences, the solution approach involves using weakly labeled data to train an initial version of the entity recognition model, then using the model to select low-confidence results from the training set for correction, and finally fine-tuning the base model with corrected data.

The main contributions are: (1) This study addresses agricultural domain entity recognition problems. To solve the lack of annotated entity recognition data in agriculture, it proposes crawling the open-source Internet database "Hudong Baike" to construct an agricultural knowledge graph and implements weak labeling of entity recognition data through distant supervision. (2) To address the problems of poor recognition performance and reliance on manual features in previous studies, this study combines agricultural entity recognition characteristics and proposes a BERT-ALA+BiLSTM+CRF model based on an attention layer aggregation mechanism, achieving optimal results on three datasets and validating the effectiveness of the layer aggregation mechanism. The purpose of this study is primarily to apply this model to agricultural entity recognition. (3) To address the high prediction latency of BERT-based models, this study uses the BiLSTM+CRF model as a student model to distill the BERT-ALA+BiLSTM+CRF model, significantly reducing the time and space complexity of the online model and making the trained model applicable on mobile devices.

The entity recognition method proposed in this study can be extended to other vertical domain entity recognition scenarios with scarce annotated data, such as medicine, education, and military domains.

We thank Researcher He Chaoxing from the Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, for providing valuable input to this study.

References

[1] COWIE J, LEHNERT W. Information extraction[J]. Communications of the ACM, 1996, 39(1): 80-91.
[2] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition[C]// The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, USA: Association for Computational Linguistics, 2016: ID N16-1030.
[3] 李贯峰, 张鹏. 一个基于农业本体的 Web 知识抽取模型 [J]. 江苏农业科学, 2018, 46(4): 201-205.
[4] 王春雨, 王芳. 基于条件随机场的农业命名实体识别研究 [J]. 河北农业大学学报, 2014, 37(1): 132-135.
[5] TSENG H, CHANG P-C, ANDREW G, et al. A conditional random field word segmenter for sighan bakeoff 2005[C]// Proceedings of the fourth SIGHAN workshop on Chinese language Processing. San Diego, USA: Association for Computational Linguistics, 2005.
[6] MALARKODI C, LEX E, DEVI S L J. Named entity recognition for the agricultural domain[J]. Research in Computing Science, 2016, 117(1): 121-132.
[7] 刘晓俊. 面向农业领域的命名实体识别研究 [D]. 合肥: 安徽农业大学, 2019.
[8] BISWAS P, SHARAN A, VERMA S. Named entity recognition for agriculture domain using word net[J]. International Journal of Computer & Mathematical Sciences, 2016, 5(10): 29-36.
[9] MILLER G A. WordNet: An electronic lexical database[M]. Massachusetts: MIT press, 1998.
[10] LI J, SUN A, HAN J, et al. A survey on deep learning for named entity recognition[J]. IEEE Transactions on Knowledge and Data Engineering, 2020 (99): 1.
[11] MINTZ M, BILLS S, SNOW R, et al. Distant supervision for relation extraction without labeled data[C]// Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. San Diego, USA: Association for Computational Linguistics, 2009: 1003-1011.
[12] ZENG D, LIU K, CHEN Y, et al. Distant supervision for relation extraction via piecewise convolutional neural networks[C]// Proceedings of the 2015 conference on empirical methods in natural language processing. Lisbon, Portugal: Association for Computational Linguistics, 2015: 1753-1762.
[13] DEVLIN J, CHANG M-W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, USA: Association for Computational Linguistics, 2018.
[14] POLINO A, PASCANU R, ALISTARH D. Model compression via distillation and quantization[EB/OL]. 2018. arXiv:1802.05668.
[15] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[EB/OL]. 2015. arXiv:1508.01991.

[16] ZHOU Z. A brief introduction to weakly supervised learning[J]. National Science Review, 2018, 5(1): 44-53.
[17] 米嘉. 大规模中文文本检索中的高性能索引研究 [D]. 北京: 中国科学院, 2005.
[18] LUO L, YANG Z, YANG P, et al. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition[J]. Bioinformatics, 2018, 34(8): 1381-1388.
[19] SOUZA F, NOGUEIRA R, LOTUFO R. Portuguese named entity recognition using BERT-CRF[EB/OL]. 2019. arXiv:1909.10649.
[20] GREFF K, SRIVASTAVA R K, KOUTNÍK J, et al. LSTM: A search space odyssey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2016, 28(10): 2222-2232.
[21] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// NIPS' 17: Proceedings of the 31st International Conference on Neural Information Processing Systems. New York, US: Curran Associates Inc., 2017: 6000-6010.
[22] DONG C, ZHANG J, ZONG C, et al. Character-based LSTM-CRF with radical-level features for Chinese named entity recognition[C]// International Conference on Computer Processing of Oriental Languages. Berlin, German: Springer, 2016: 239-250.
[23] YAN H, DENG B, LI X, et al. Tener: Adapting transformer encoder for name entity recognition[EB/OL]. 2019. arXiv:1911.04474.
[24] JIANG S, ZHAO S, HOU K, et al. A BERT-BiLSTM-CRF model for chinese electronic medical records named entity recognition[C]// 2019 12th International Conference on Intelligent Computation Technology and Automation. Piscataway, New York, USA: IEEE, 2019: 166-169.
[25] JAWAHAR G, SAGOT B, SEDDAH D. What does BERT learn about the structure of language?[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. San Diego, USA: Association for Computational Linguistics, 2019.
[26] OPITZ J, BURST S. Macro F1 and Macro F1[EB/OL]. 2019. arXiv:1911.03347.
[27] GRAVE E, BOJANOWSKI P, GUPTA P, et al. Learning word vectors for 157 languages[C]// Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA), 2018.
[28] KINGMA D P, BA J J A P A. Adam: A method for stochastic optimization[EB/OL]// 3rd International Conference on Learning Representations. Ithaca, NY: arXiv.org, 2015: 13.

Note: Figure translations are in progress. See original paper for figures.