Journal of Animal & Plant Sciences, 36(1): 2026, Page: 213-226
ISSN (print): 1018-7081; ISSN (online): 2309-8694
https://doi.org/10.36899/JAPS.2026.1.0018

CHINESE-LANGUAGE ENTITY RELATION EXTRACTION FOR WHEAT DISEASES AND PESTS BASED ON REMOTE SUPERVISION AND BIDIRECTIONAL TRANSFORMERS

Y. Che [1,3], S. Xiong [2,3], S. Xi [4], D. Zhang [1], X. Ma [2,3] and L. Xi [1,3,*]

[1]College of Information and Management Science, Henan Agricultural University, Zhengzhou 450046, China
[2]College of Agronomy, Henan Agricultural University, Zhengzhou 450046, China
[3]Henan Engineering Laboratory for Farmland Environmental Monitoring and Control Technology, Zhengzhou 450046, China
[4]University of London, London, WC1E 7HU, UK
*Corresponding author's email: 13111560238@163.com

ABSTRACT

In the realm of wheat disease and pest management, extracting domain-specific knowledge presents a formidable challenge. This challenge is further amplified by the lack of publicly available Chinese entity-relation extraction datasets and the exorbitant costs associated with manual annotation stemming from the specialized nature of the domain. In response to these challenges, we utilized remote supervision to match relevant triplets from CN-DBpedia and Ownthink knowledge bases with unstructured texts, followed by manual correction to construct WheatCRE, a Chinese entity relation extraction dataset for wheat diseases and pests. The WheatCRE dataset comprises 1,681 labeled samples covering six relationship categories: Distribution Range, Alias, Damage Parts, Damage Crops, Genus Orders, and Genus Families. Subsequently, we proposed a novel model called BE-CRE (BERT-Entity Chinese Relation Extraction), which combines Bidirectional Encoder Representations from Transformers with entity representations. The model uses BERT to obtain dynamic character representations and target entity representations, adopting concatenation method to fuse features. By making full use of the implicit meanings of entities, the model can obtain more accurate features. Comprehensive experiments were conducted on the WheatCRE dataset comparing different optimization algorithms, training parameters, relation extraction models, and pre-training models. Our proposed BE-CRE model achieved superior performance compared to baseline models including BiLSTM-Attention, BiGRU-Attention, and BERT-Softmax, with Precision-Macro (P-M), Recall-Macro (R-M), and F1-Macro (F1-M) values of 89.37%, 89.4%, and 89.29% respectively. Furthermore, we conducted comparative experiments on a public Chinese entity relation extraction dataset (CharacterCRE) to evaluate the generalization ability of our model, achieving the best F1-Macro value of 78.31%. These results demonstrate the effectiveness and applicability of our proposed model in wheat disease and pest relation extraction. BE-CRE distinguishes itself from existing models by integrating BERT with entity representations and using remote supervision to construct a specialized dataset, enabling more accurate and context-aware entity relation extraction for agricultural applications.

Key words: Remote supervision, Chinese entity relation extraction, Wheat diseases and pests, BERT.

Published first online December 05, 2025 Published final January 20, 2026

INTRODUCTION

Wheat (*Triticum aestivum* L.) diseases and pests represent significant threats to global food security, causing annual production losses of 20-30% worldwide (Jalli *et al.*, 2021; Hussain *et al*., 2024). These losses translate to billions of dollars in economic impact annually, affecting agricultural producers and food availability globally. Among various wheat diseases, rust pathogens are particularly destructive, with modern detection methods utilizing unmanned aerial vehicles and visual question answering systems achieving promising results (Joshi *et al.*, 2024; Nanavaty *et al.*, 2024). Beyond direct yield losses, wheat diseases and pests compromise grain quality, reduce nutritional content, and increase production costs through intensive pesticide applications. Early detection and accurate identification of plant diseases are crucial for implementing effective control measures (Hossain and Deb, 2021).

Entity relation extraction serves as a fundamental component in knowledge graph construction and plays a crucial role in organizing scattered agricultural information (Li *et al*., 2020a; Wang and Zhao, 2024). Recent advances have demonstrated the

effectiveness of pre-trained language models combined with graph neural networks in Chinese geological domains, achieving superior performance in entity relation tasks (Lv *et al*., 2022). Given the critical importance of wheat in global agriculture, effective knowledge management and prevention strategies require sophisticated information processing technologies that can handle the complexity and domain-specific nature of agricultural knowledge.

Entity relation extraction is essential for knowledge graph construction and significantly influences knowledge representation quality (Han and Wang, 2024). With the rapid advancement of deep learning, deep neural network-based methods for Chinese entity relation extraction have become prevalent across various fields (Zhao *et al*., 2023). However, in specialized domains like wheat diseases and pests, challenges such as limited labeled datasets, scattered information, complex text structures, varied expression patterns, and extensive domain-specific terminology make generic relation extraction methods inadequate (Yao *et al*., 2024).

Contemporary approaches integrate advanced machine learning methods, with modern relation extraction methodologies embracing novel decomposition strategies that separate tasks into interconnected subtasks, effectively capturing correlations while reducing noisy entity impacts (Yu *et al.*, 2020). Enhanced models combining attention mechanisms with pointer annotation have demonstrated improved performance, particularly in addressing overlapping relations (Yu *et al*., 2020). Graph-based approaches have emerged as powerful alternatives, with graph convolutional neural networks and weighted diffusion graph methods successfully applied to Chinese entity relation extraction tasks (Chen *et al.*, 2024). However, the application of these advanced techniques to Chinese agricultural texts presents unique challenges, including complex character structures, varied expression patterns, and extensive domain-specific terminology.

Agricultural applications have witnessed significant progress through specialized attention mechanisms with external knowledge reinforcement and distant supervision techniques (Li *et al.*, 2020b; Yue *et al.*, 2020). Joint extraction algorithms have been developed specifically for agricultural domains, addressing complex relationships between diseases, pests, weeds, and chemical agents (Yue *et al*., 2020). Domain-specific datasets have been created for various crops (Yao *et al.*, 2024). BERT-based approaches have shown exceptional effectiveness in agricultural knowledge graph construction and biomedical literature analysis, with successful implementations in crop disease and pest knowledge extraction (Wu *et al*., 2020) and long biomedical text processing (Li *et al*., 2024), though attention mechanisms adapted for agricultural applications often focus on attention weights while overlooking entity-specific semantic information (Zhang *et al*., 2021). Contemporary developments have achieved remarkable results in crop disease detection using machine learning approaches (Dolatabadian *et al*., 2025), with comprehensive reviews highlighting the effectiveness of deep learning techniques for plant disease identification (Pacal *et al*., 2024), and deep learning algorithms for wheat leaf disease identification (Xu *et al*., 2023; Long *et al.*, 2023). Despite these advances, the scarcity of high-quality, annotated Chinese datasets in agricultural domains limits the development and evaluation of specialized models.

Existing approaches face significant limitations when applied to wheat diseases and pests, particularly in overlooking entity-specific semantic information while emphasizing attention mechanisms, resulting in incomplete utilization of contextual knowledge. Current methods often treat all entities equally without considering their specific semantic properties and domain-specific meanings, leading to suboptimal performance in specialized agricultural contexts. This research addresses these limitations through two key innovations. First, the study introduces a specialized remote supervision framework designed specifically for wheat diseases and pests, leveraging external knowledge bases CN-DBpedia and Ownthink while incorporating systematic manual correction to create WheatCRE, a comprehensive Chinese entity relation extraction dataset. Second, the research proposes BE-CRE, a novel model architecture that explicitly integrates BERT with entity representations, departing from conventional attention-only approaches by incorporating entity-specific information as additional constraints to enhance semantic understanding and relation classification accuracy. This study contributes to agricultural knowledge management by developing domain-specific tools for wheat diseases and pests' information extraction, providing technical foundations for intelligent agricultural systems and crop protection strategies.

MATERIALS AND METHODS

RE Dataset Construction Based on Remote Supervision: Remote supervision effectively addresses the shortage of annotation samples by leveraging entity-relationship pairs from existing knowledge bases for automatic labeling in self-constructed corpora (Yue *et al*., 2020). This approach enables systematic annotation through the assignment of relationship labels based on external knowledge sources. In the domain of wheat diseases and pests in China, the scarcity of publicly available relation extraction datasets and the high cost of manual annotation due to domain specialization present significant challenges. To overcome these obstacles, this research introduces WheatCRE, a Chinese entity-relation extraction dataset for wheat diseases and pests,

combining remote supervision methodology with systematic manual correction processes.

The remote supervision framework operates through a systematic entity-relationship mapping process that establishes connections between structured knowledge base triplets and unstructured textual content, as illustrated in Figure 1. The annotation process encompasses three interconnected stages: preprocessing the raw corpus and generating alternative sentences alongside the wheat diseases and pest's domain dictionary (WpdDict), extracting wheat diseases and pests related alternative triples from knowledge bases based on WpdDict, and performing matching and alignment operations. This systematic approach ensures comprehensive coverage while maintaining annotation quality through structured workflow management.

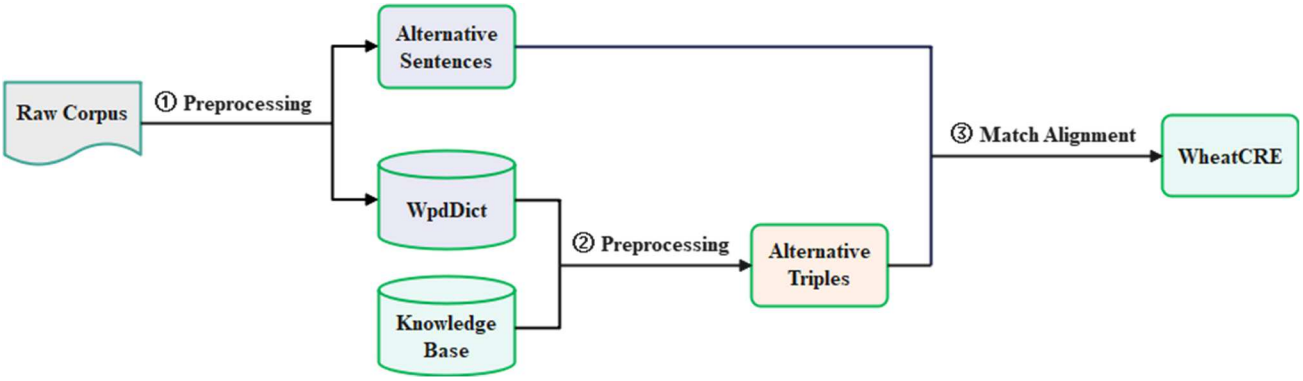


Fig. 1. Annotation process based on remote supervision for WheatCRE dataset construction. The framework consists of three stages: (1) Chinese corpus collection from authoritative sources and alternative sentence generation, producing 7,364 sentences and WpdDict with 4,125 entities; (2) Triple extraction from CN-DBpedia and Ownthink knowledge bases with quality enhancement through normalization and de-duplication; (3) Match alignment between triplets and sentences to generate 1,681 labeled samples in format <Head Entity, Tail Entity, Relationship, Sentence>.

Chinese Corpus Collection and Alternative Sentences Generation: With the rapid advancement of agricultural informatization, substantial unstructured data on wheat diseases and pests has been accumulated across various digital platforms. Authoritative sources including China Crop Germplasm Information Network (https://www.cgris.net) and Baidu Wikipedia (https://baike.baidu.com) serve as primary data repositories, supplemented by specialized publications such as "Atlas of Diagnosis and Control of Wheat Diseases and Pests" and "Identification and Control of Wheat Diseases and Pests." These sources provide comprehensive coverage of wheat disease and pest information, encompassing over 60 diseases and 40 insect pests with their associated relationships.

The corpus preprocessing involves multiple stages of text standardization and quality enhancement. Text format standardization includes digitization of printed materials and conversion to uniform TXT format. Content filtering removes irrelevant elements such as garbled codes, URLs, headings, and special symbols while preserving semantic integrity. Sentence completion addresses fragmented information by adding missing subjects based on contextual analysis. Through systematic preprocessing guided by punctuation marks such as "。", "？", "！", and "...", 7,364 alternative sentences are generated. During preprocessing, expert guidance facilitates the extraction of public entities and their synonyms, leading to the construction of WpdDict, which includes 4,125 entities related to wheat diseases and pests.

Triple Extraction and Quality Enhancement; The methodology leverages two comprehensive knowledge bases to extract domain-relevant triplets for remote supervision. CN-DBpedia, developed and maintained by Fudan University's Knowledge Workshop Lab, encompasses over 9 million entities and 67 million entity-relationship triplets in the standardized format <Entity, Relationship, Entity>. Ownthink represents the largest open-source Chinese knowledge graph, containing 140 million triplets in both <Entity, Relationship, Entity> and <Entity, Attribute, Attribute Value> formats, offering extensive coverage of domain-specific information.

The triple extraction process utilizes WpdDict as a filtering mechanism to identify relevant triplets from the extensive knowledge bases. The extraction algorithm systematically queries both CN-DBpedia and Ownthink using WpdDict entities as search keys, retrieving all triplets containing these entities as either head or tail components. This targeted approach ensures that extracted triplets maintain relevance to the wheat diseases and pests domain while leveraging the comprehensive coverage provided by large-scale knowledge bases.

Analysis of extracted triplets revealed several systematic issues requiring resolution: multiple

descriptions exist for identical relationship categories, certain triplets contain more than one entity in tail positions, and duplicate triplets are present throughout the dataset. To address these challenges, relationship normalization procedures consolidate diverse terms such as "Damage Crops," "Mainly Grazing Damage," "Endanger Crops," "Mainly Damage Crops," and "Hosts" under the standardized relationship "Damage Crops." Complex triplets containing multiple entities undergo manual correction and splitting processes, as demonstrated in Fig. 2.

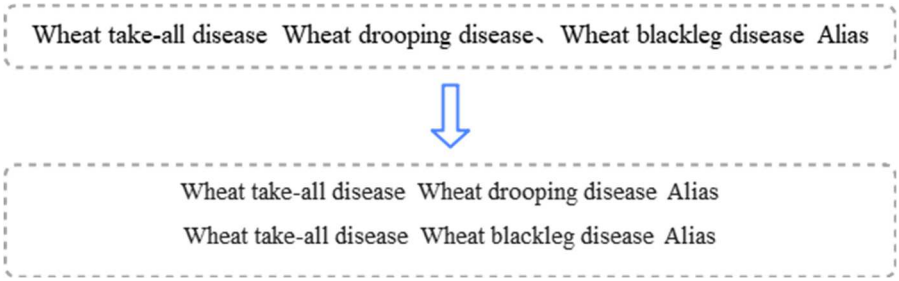


Fig. 2. Example of entity splitting process for complex triplets containing multiple tail entities. The original triplet with multiple geographical locations is manually split into separate one-to-one relationships to ensure data quality and prevent confusion during model training. This preprocessing step is essential for maintaining clear entity-relationship mappings in the WheatCRE dataset.

De-duplication procedures eliminate redundant entries, and the distribution of various relationship categories is analyzed as shown in Figure 3. To ensure dataset quality, the top six relationship categories are selected: "Distribution Range," "Alias," "Damage Parts," "Damage Crops," "Genus Orders," and "Genus Families," resulting in 2,143 alternative triplets in the format <Head Entity, Tail Entity, Relationship>.

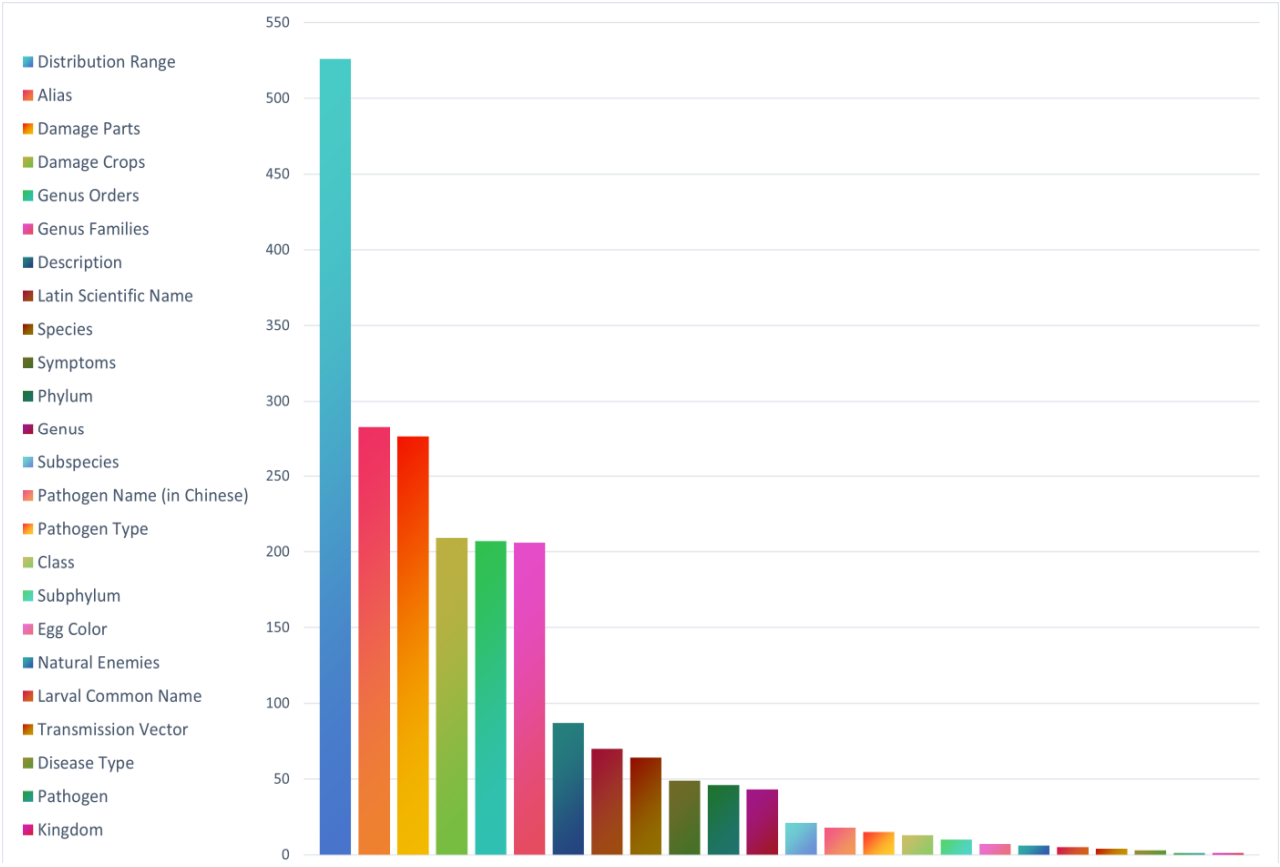


Fig. 3. Distribution of Relation Categories

Match Alignment and Sample Generation: The alternative triplets with relational labels undergo systematic matching within the self-developed corpus. The matching process identifies textual instances containing both head and tail entities from each triplet, subsequently assigning the corresponding relationship label to create annotated samples. For example, the triplet "Wheat powdery Damage parts" is matched with the sentence "Wheat powdery damages leaves" to generate the labeled sample format <Head Entity, Tail Entity, Relationship, Sentence>. Through this systematic matching process, 1,681 samples are obtained. The reduction from 2,143 alternative triplets to 1,681 labeled samples occurred because not all triplets could be successfully matched with corresponding sentences in the corpus. Only those triplets where both the head entity and tail entity appeared together in at least one sentence from the alternative corpus were converted into labeled samples. Specific labeling examples are illustrated in Figure 4.

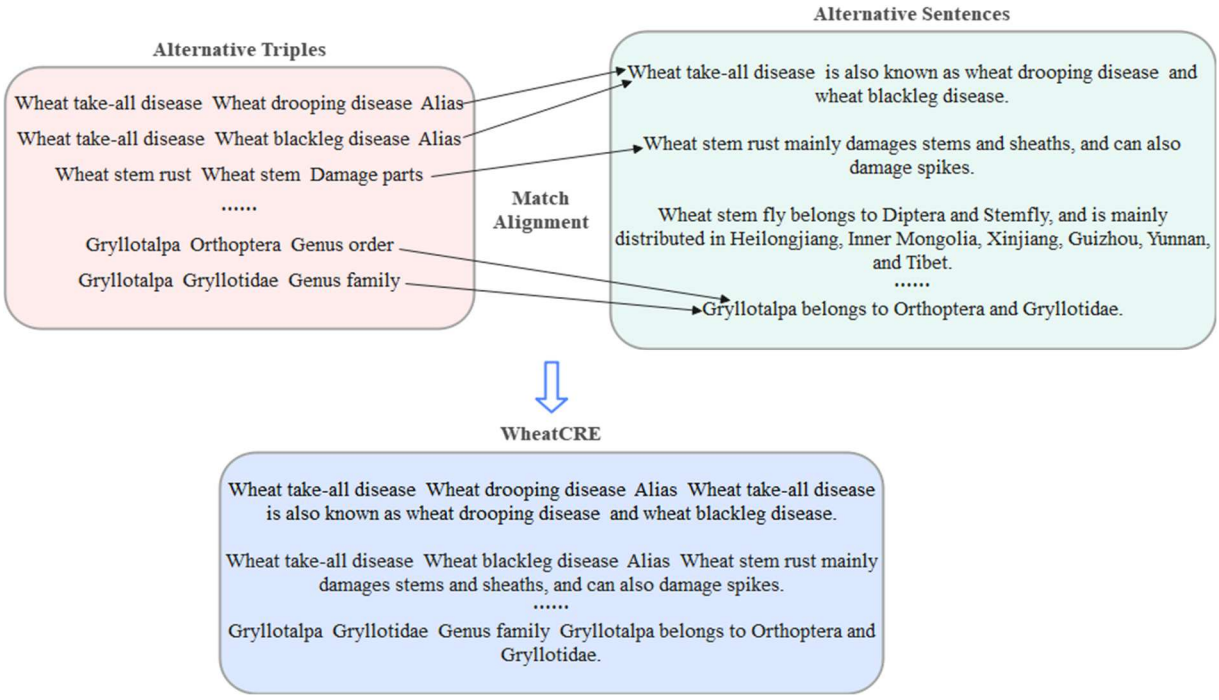


Fig. 4. Example of annotation samples showing the match alignment process between alternative triplets and alternative sentences to generate WheatCRE dataset. The process demonstrates how triplets containing entity relationships (such as 'Wheat take-all disease', 'Wheat drooping disease', 'Alias') are matched with corresponding sentences from the corpus to create labeled samples for the Chinese entity relation extraction dataset.

Considering the potential impact of uneven relationship category distribution on extraction performance, this research implements stratified sampling to ensure balanced representation. For each relationship category, 80% of samples are randomly selected for the training set while 20% comprise the test set, as detailed in Table 1.

BE-CRE Model Architecture and Innovation: The BE-CRE model introduces a novel architecture that explicitly integrates BERT-based contextual representations with entity-specific information, addressing limitations in existing approaches that rely solely on attention mechanisms. The model architecture consists of three interconnected components: a BERT encoding layer for contextual representation learning, a feature fusion layer for integrating multiple information sources, and a classification layer for relationship prediction, as illustrated in Figure 5. This design enables the model to leverage both global contextual information and local entity-specific features, resulting in enhanced semantic understanding and improved classification accuracy.

Table 1. Distribution of WheatCRE Dataset

ID	Relationship Categories	Training Set	Test Set
0	Distribution Range	418	105
1	Alias	226	57
2	Damage Parts	213	54
3	Damage Crops	163	42
4	Genus Families	163	41
5	Genus Orders	159	40

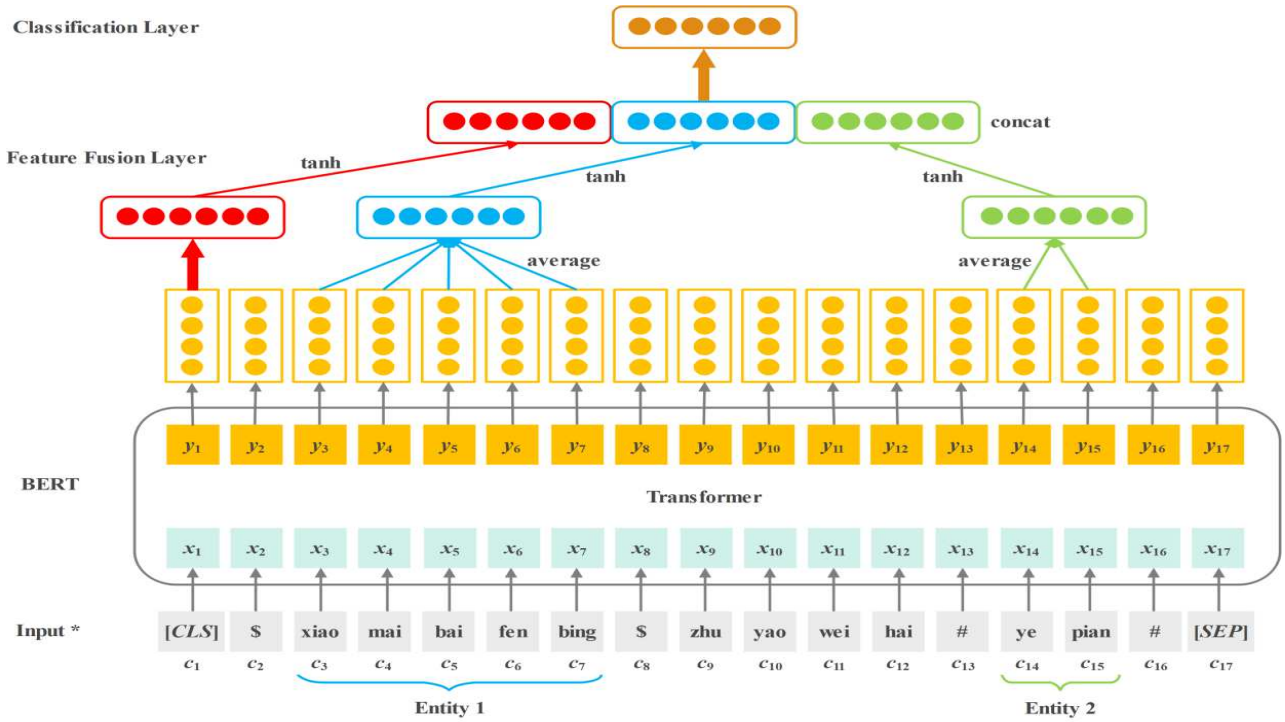


Fig. 5. BE-CRE Model Framework (*: "xiao", "mai", "bai", "fen", "bing" are the pinyin of the Chinese characters, as are "zhu", "yao", "wei", "hai", and "ye", "pian", which correspond to "wheat powdery mildew", "mainly damage", "leaves", respectively.)

BERT Layer Implementation BERT: (Bidirectional Encoder Representations from Transformers) serves as the foundational encoding component, utilizing bidirectional transformer architecture to automatically learn contextual features and generate dynamic character vector representations. The model excels in various NLP applications through fine-tuning mechanisms that enhance downstream task performance. The input representation combines three embedding features: Character Embedding, Segment Embedding, and Position Embedding, enabling comprehensive semantic understanding.

For a given input sequence, '[CLS]' and '[SEP]' are added at the beginning and end of the sentence, respectively, while '$' and '#' to mark the position of entity e1 and entity e2 in the sentence. As shown in Figure 6, after inserting the marker in "Wheat powdery mildew mainly damages leaves", the input becomes "[CLS]$Wheat powdery mildew$mainly damages#leaves#[SEP]", where $\{c_3, c_4, .., c_7\}$ and $\{c_{14}, c_{15}\}$ denote entities e1 and e2, respectively.

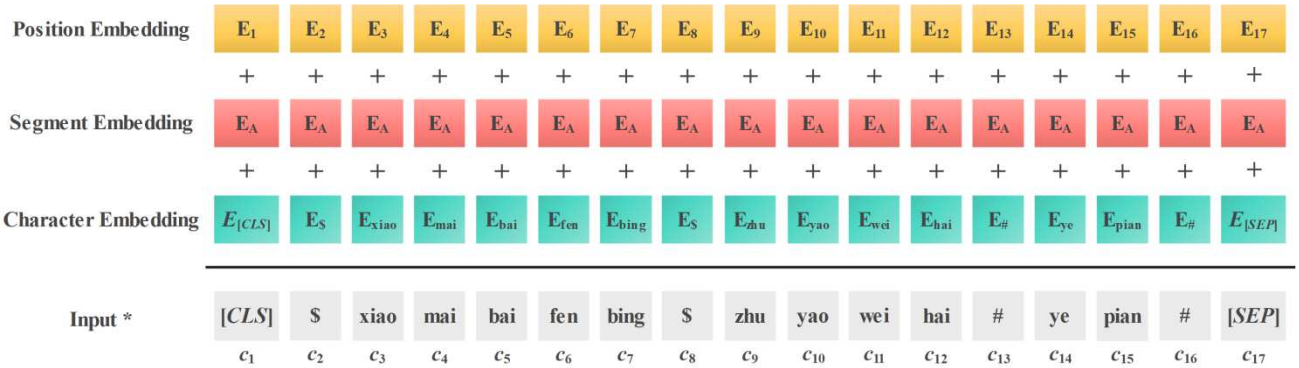


Fig. 6. Input Representation of BERT Model (*: "xiao", "mai", "bai", "fen", "bing" are the pinyin of the Chinese characters, as are "zhu", "yao", "wei", "hai", and "ye", "pian", which correspond to "wheat powdery mildew", "mainly damage", "leaves", respectively.)

representations from texts related to wheat diseases and pests (Devlin *et al.*, 2019). The process of feature extraction based on BERT is visually depicted in Figure 7

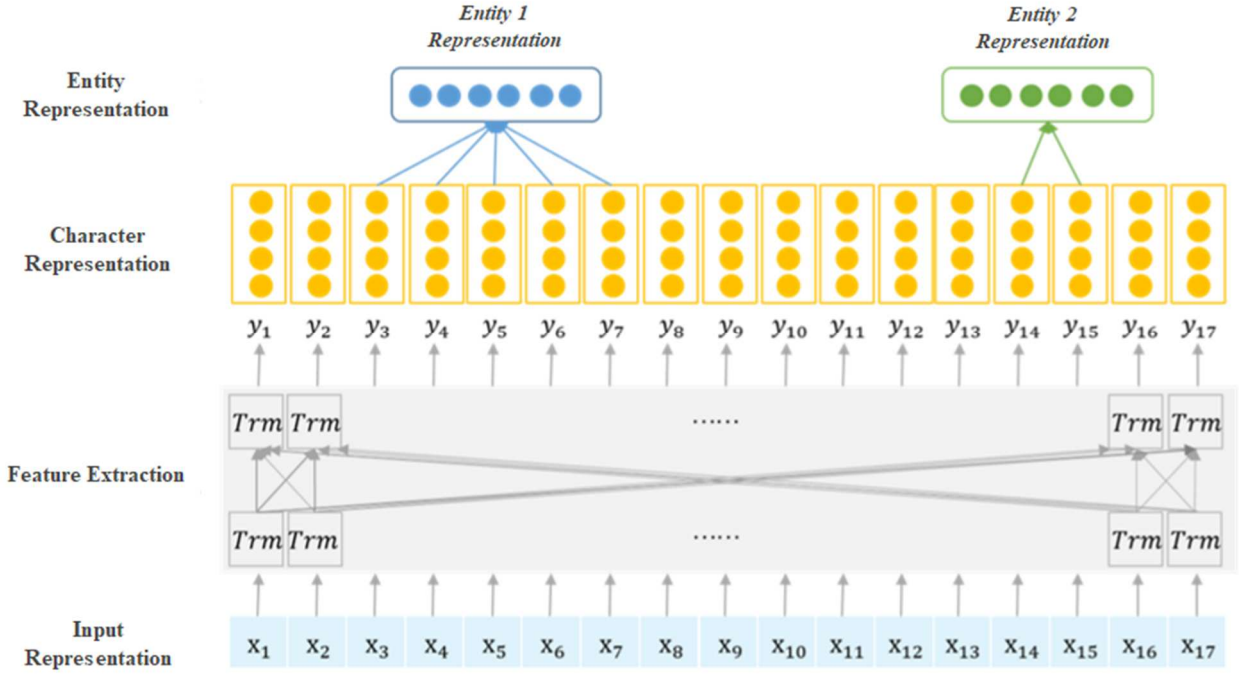


Fig. 7. BERT-based feature extraction process generating sentence and entity representations. Input tokens are processed through Transformer encoders to produce character-level representations H=[h1, h2, ..., hm]. Entity representations are obtained by averaging character vectors within entity boundaries, capturing both global sentence context and local entity semantics for subsequent fusion and classification.

Given a sentence $S=\{c_1,c_2,\ldots,c_m\}$ of length m, the two entities in the sentence are $e1=\{c_a,c_{a+1},\ldots,c_b\}$, and $e2=\{c_u,c_{u+1},\ldots,c_v\}$ with c_i representing the i_{th} character, a and b represent the start position and end position of entity e1, and u and v represent the start position and end position of entity e2. The input representation of the character is obtained by summing the three embedding vectors mentioned above, denoted as x_i, which is then fed to the Transformer encoder for feature extraction and the final output is y_i.

Denote the character vector representation of the sentence as Es, as shown in Equation (1). Perform the sum and average operation on the character vectors constituting entities e1 and e2 to obtain the vector representations of the two entities, denoted as E_{e1} and E_{e2}, as shown in Equation (2) and (3), respectively.

$$E_S=y_1,y_2,\ldots,y_m \quad (1)$$

The core of the BERT model is multiple Transformer encoders crafted to acquire dynamic character vector representations and target entity vector

$$E_{e1}=\frac{1}{b-a+1}\sum_{i=a}^{b}y_i \quad (2)$$

$$E_{e2}=\frac{1}{v-u+1}\sum_{i=u}^{v}y_i \quad (3)$$

Feature Fusion Layer: In the process of extracting specific entity relationships, utilizing entity information serves as a form of constraint (Shi and Lin, 2019). For example, " wheat powdery mildew" and " leaves" are known, and after model training, the less relevant relationships such as "distribution region" and "genus order" can be removed to certain extent. Utilizing the specific meanings implied by entity information of wheat diseases and pests to provide additional effective information enables the model to learn more accurate features, which is conducive to judging entity relationship categories.

After obtaining the dynamic character representations and target entity representations through BERT, feature fusion is performed by splicing, and entity representations are utilized to strengthen the model training, so that the output features can express richer semantic information, and thus enhance the Chinese

entity relationship extraction results for wheat diseases and pests (Shim and Shim, 2023).The character vector representation E_s and entity representations E_{e1} and E_{e2} are subjected to feature fusion through Equation (4).

$$H = concat(E_s, E_{e1}, E_{e2}) \quad (4)$$

Classification Layer: In this study, the Softmax classifier is used to normalize the probability of the relationship categories, and the relationship classification is realized by comparing the probability. The output H of the feature fusion layer is used as the input of the classification layer, and for a given input sample S, the probability of each relationship category is calculated by Softmax, as shown in Equation (5). And output the one with the highest probability among the predefined six relationship categories to achieve the purpose of relationship extraction, as shown in Equation (6).

$$P(R \mid S) = softmax(WH + b) \quad (5)$$

$$\hat{R} = argmaxP(R \mid S) \quad (6)$$

where W is the weight matrix, b is the bias vector, R is the set of relationship categories, P is the predicted probability of each relationship category, and $\hat{R}$ is the predicted relationship label. In Figure 6, the relationship $\hat{r}$ between "wheat powdery mildew" and "leaves" predicted by the BE-CRE model is "damage part".

In multi-classification tasks, the cross-entropy loss function is usually used as the loss function, as shown in Equation (7).

$$J(\theta) = \sum_{i=1}^{T} logp(R^i \mid S^i, \theta) \quad (7)$$

where T denotes the total number of training samples (S^i, θ) and θ denotes all parameters in the model.

Feature Fusion and Classification: The feature fusion mechanism represents a key innovation of the BE-CRE model, enabling systematic integration of sentence-level and entity-level representations. Unlike traditional approaches that rely solely on attention weights, this fusion strategy explicitly incorporates entity-specific information as additional constraints for relationship classification. The concatenation-based fusion operation $H = concat(E_s, E_{e1}, E_{e2})$ produces enriched feature vectors that capture both contextual dependencies and entity-specific semantic properties, enabling the model to utilize entity information as semantic constraints for filtering irrelevant relationship categories.

The classification layer employs a Softmax classifier to compute relationship probabilities through $P(R|S) = softmax(WH + b)$, where W represents the weight matrix and b denotes the bias vector. The predicted relationship $\hat{R} = argmax\ P(R|S)$ determines the final classification result. The optimization process utilizes cross-entropy loss function $J(\theta) = \sum_{i=1}^{T} logp(R^i \mid S^i, \theta)$ to minimize prediction errors, where T represents the total number of training samples and θ denotes all model parameters.

RESULTS

Experimental Environment: The experiments were conducted in Windows 10, and the GPU was NVIDIA Tesla V100-PCIE-16GB-LS. a Pytorch version of BERT-base-Chinese was used, containing a 12-layer Transformer, a hidden-layer dimension of 768, and a 12-head attention mechanism. The specific parameters were shown in Table 2.

In order to verify the generalization of the BE-CRE model, in addition to conducting comparative experiments on the self-built WheatCRE dataset, comparative experiments were conducted on the public Chinese Character Relationship Extraction dataset (CharacterCRE). The CharacterCRE is a human entity-relation dataset, and consists of 12 relationship categories with a total of 200,000 labeled samples, from which 10,000 samples were selected. For WheatCRE and CharacterCRE, the training set and test set were divided according to the ratio of 8:2, respectively.

Table 2. Model Hyperparameters Configuration

Parameters	Values
Hidden_dim	768
Learning_rate	0.00002
Max_len	128
Epoch	100
Batch_size	32
Pos_dim	50

Note:Hidden_dim: hidden layer dimension; Max_len: maximum sequence length; Pos_dim: position embedding dimension

Evaluation Indicators: Macro Precision (P-M), Macro Recall (R-M) and Macro F1-score (F1-M) were selected as the evaluation Indicators of the model, where F1-M is the comprehensive evaluation indicators of P-M and R-M with the following equations.

$$P = \frac{TP}{TP + FP} \times 100\% \quad (8)$$

$$R = \frac{TP}{TP + FN} \times 100\% \quad (9)$$

$$F1 = \frac{2 \times P \times R}{P + R} \times 100\% \quad (10)$$

$$\text{P-M} = \frac{1}{n}\sum_{i=1}^{n} \text{P}_i \qquad (11)$$

$$\text{R-M} = \frac{1}{n}\sum_{i=1}^{n} \text{R}_i \qquad (12)$$

$$\text{F1-M} = \frac{1}{n}\sum_{i=1}^{n} \text{F1}_i \qquad (13)$$

Where TP denotes the number of positive samples with correct prediction, FP denotes the number of positive samples with incorrect prediction, FN denotes the number of negative samples with incorrect prediction, and n denotes the number of predefined entity-relationship categories for wheat diseases and pests.

Relationship Extraction Results of BE-CRE Model

Impact of Different Optimization Algorithms on the Results

Gradient descent optimizer (GDO) is an optimization algorithm, which is widely used in deep learning to obtain the minimized loss function and optimal parameter values. The commonly used GDO algorithms include Stochastic gradient descent (SGD), Resilient Back Propagation (Rprop), Adaptive Delta (Adadelta), Adaptive moment estimation (ADAM), etc.

For the entity-relationship extraction model BE-CRE, the learning rate was set to 0.00002, the dropout rate was 0.5 and the batch size was 32. The extraction results of the four optimization algorithms on the BE-CRE model were shown in Figure 8.

As could be seen from Figure 8, in the process of model training, when using the two optimization algorithms of SGD and Adadelta, was lower than 70%, which is worse compared to the other two optimization algorithms; when using the Rprop optimization algorithm, the final F1-M value is 85.54%, which is lower than that of ADAM algorithm; and when using the ADAM optimization algorithm, the F1-M achieved the highest value of 89.29%, which is 3.75% higher than Rprop; in addition, ADAM has the advantages of occupying fewer resources and faster convergence of the model, so ADAM algorithm is chosen in this paper.

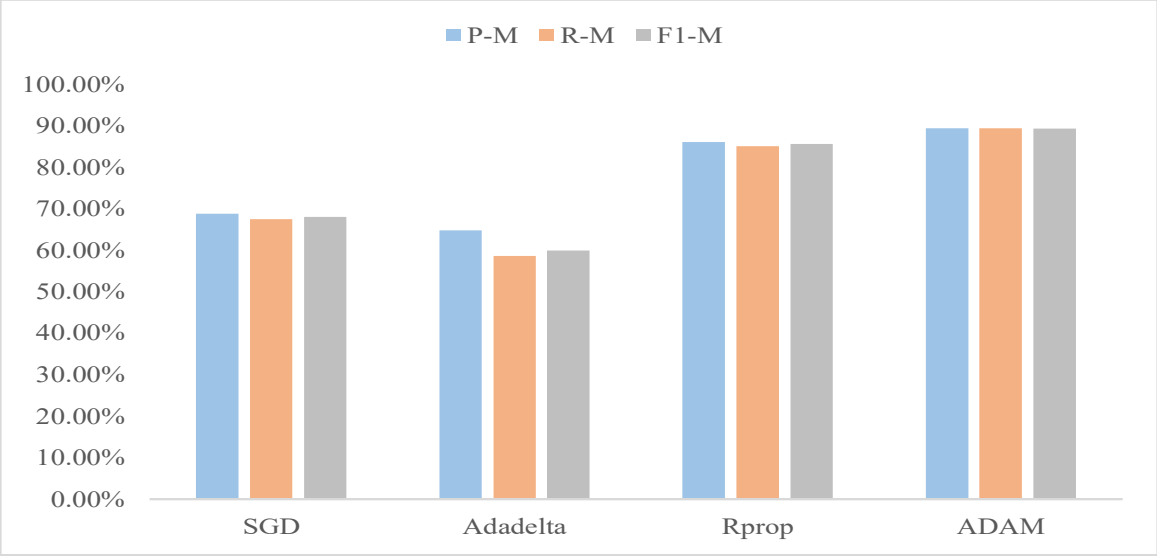
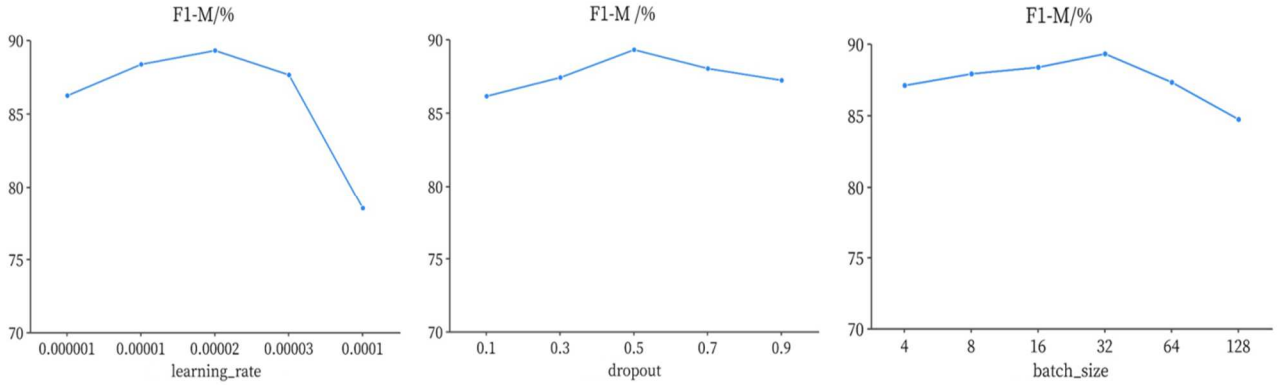


Fig. 8. Performance comparison of four optimization algorithms (SGD, Adadelta, Rprop, and ADAM) on the BE-CRE model using WheatCRE dataset. The figure shows Precision-Macro (P-M), Recall-Macro (R-M), and F1-Macro (F1-M) scores for Chinese entity relation extraction in wheat diseases and pests domain. ADAM optimizer achieved the highest F1-M score of 89.29%, demonstrating superior convergence and performance compared to other algorithms. Learning rate was set to 0.00002, dropout rate to 0.5, and batch size to 32 for all experiments.

Impact of Different Parameters on the Results: This study employs a step-by-step search strategy, fixing two parameters at a time and gradually adjusting the third to determine the optimal value. Based on existing literature and general experience, the search spaces for the learning rate, dropout rate, and batch size were defined as [0.000001, 0.0001], [0.1, 0.9], and [4, 128], respectively. Using the Adam optimizer, the experiment began by adjusting only the learning rate to observe its impact on model performance. Next, the dropout rate was varied to assess its effect, followed by adjustments to the batch size to evaluate its contribution to overall model performance. The experiment revealed that the best performance for Chinese entity-relationship extraction in the context of

wheat diseases and pests was achieved when the learning rate, dropout rate, and batch size were set to 0.00002, 0.5, and 32, respectively. The effects of these three parameters on the F1-M score were shown in Figure 9.



(a) **(b)** **(c)**

Fig. 9. Impact of hyperparameters on BE-CRE model performance using F1-Macro scores. (a) Learning rate optimization across range [0.000001, 0.0001] with optimal value 0.00002; (b) Dropout rate analysis from [0.1, 0.9] with best performance at 0.5; (c) Batch size evaluation from [4, 128] showing optimal value 32. All experiments used ADAM optimizer with systematic parameter search strategy.

Comparison Results and Analysis on WheatCRE Dataset:

Extraction Results for Different RE Models: In this paper, 4 models, BiLSTM-ATT, BiGRU-ATT, BiLSTM-2ATT, and BiGRU-2ATT, are selected for comparison, in which the word-level attention mechanism is integrated in the first two models, respectively, while the latter two models are based on the first two models, respectively, and add another layer of sentence-level attention mechanism. The relationship extraction results were shown in Table 3.

Table 3. Performance comparison of different relation extraction models on WheatCRE dataset

Models	P-M/%	R-M/%	F1-M/%
BiLSTM-ATT	68.63	66.19	67.28
BiGRU-ATT	67.61	63.25	65.23
BiLSTM-2ATT	82.89	82.12	81.49
BiGRU-2ATT	80.23	79.58	78.94
BE-CRE	89.37	89.4	89.29

Note:P-M: Precision-Macro (%); R-M: Recall-Macro (%); F1-M: F1-Macro (%). ATT: Attention mechanism; 2ATT: Dual-attention mechanism (word-level and sentence-level).

As could be seen from Table 3, in the case of single word-level attention mechanism, the extraction results of BiLSTM-ATT and BiGRU-ATT models are poor, and the values of the three indicators, P-M, R-M, and F1-M, were all lower than 70%; after the addition of the sentence-level attention mechanism, the relationship extraction results were significantly improved, and the F1-M of BiLSTM-2ATT and BiGRU-2ATT were improved by 14.21% and 13.71% respectively. The BE-CRE model demonstrated an improvement of 9.14%, 9.82%, and 10.35% across the three evaluation indicators compared to BiLSTM-2ATT.

Extraction Results for Different Pre-trained Models: Four pre-trained models were compared on the WheatCRE dataset, and the results were shown in Table 4.

Table 4. Extraction Results of Different Pre-trained Models

Models	P-M/%	R-M/%	F1-M/%	Training time/s
ALBERT-softmax	84.88	83.73	84.21	20.45
RoBERTa-softmax	85.87	87.23	86.33	18.85
BERT-softmax	87.8	88.65	87.99	21.75
BE-CRE	89.37	89.4	89.29	23.45

Note:P-M: Precision-Macro (%); R-M: Recall-Macro (%); F1-M: F1-Macro (%). All models used ADAM optimizer.

As could be seen in Table 4, the first three relationship extraction models, although no entity representation was introduced, perform well in their extraction results, with all three evaluation indicators higher than 84%. Among them, BERT had the highest F1-M value of 87.99%, which was improved by 2.56% and 1.45% compared with ALBERT and RoBERTa respectively, but its training takes the longest time, taking 21.75s to train one epoch. Although each epoch of training the model takes 1.7s more time, were improved by 1.57%, 0.75% and 1.3% respectively compared with BERT-softmax.

Extraction Results for Different Relationship Categories: Compared with other models, the BE-CRE

model performed better in the task of extracting Chinese entity relationships for wheat diseases and pests. The extraction results of 6 categories of relationships based on the BE-CRE model were shown in Table 5.

Table 5. Results of Entity Relation Extraction for Wheat Diseases and Pests

Relationship Categories	P-M/%	R-M/%	F1-M/%
Distribution Range	94.85	92.59	93.71
Alias	87.72	89.29	88.50
Damage Parts	81.97	94.34	87.72
Damage Crops	80.77	77.78	79.25
Genus Orders	95.92	91.26	93.53
Genus Families	95.01	91.14	93.03

Note:P-M: Precision-Macro (%); R-M: Recall-Macro (%); F1-M: F1-Macro (%).

As could be seen from Table 5, Distribution Range" had the best extraction result, with a F1-M value of 93.71%, which was due to the fact that this relationship category has the largest number of training samples. The relationship categories "Genus Orders" and "Genus Families" had the lowest number of training samples, but were 93.53% and 93.03% respectively, which are only slightly lower than that of the "Distribution Range" relationship category, because the labeled samples of these two relationship categories are shorter in length, with fixed sentence patterns and less redundant information in the sentences. The number of labeled samples for the "Alias" and "Damage Parts" relationship categories was comparable, second only to the "Distribution Range" relationship category. However, their F1-M values are lower than 89%, which are 5.21% and 5.99% less than those of "Distribution Range". The "Damage Crops" relationship category has the worst extraction results, with a F1-M value of 79.25%, which was lower than 80%, due to the fact that the labeled samples of this relationship category are longer in length, with a complex sentence structure, and some of the labeled samples have entity nesting in them.

Comparison Results and Analysis on CharacterCRE Dataset: In order to verify the generalization of BE-CRE model, experiments were conducted on the Chinese Character Relationship Extraction dataset (Character CRE), and Table 6 shows the extraction results of all models.

As could be seen from Table 6, BiLSTM-ATT and BiGRU-ATT, which only have a single-level word attention mechanism, have the worst extraction results, and their F1-M values are all below 55%; the sentence-level attention mechanism was introduced into two models, BiLSTM-2ATT and BiGRU-2ATT, and the relation extraction results have been improved to a certain extent, with F1-M values of 59.81% and 56.36%, respectively; prior to the fusion of entity representations, the three pre-trained models outperform the initial four models, with their F1-M values above 70%, among which the BERT-Softmax model had the best extraction results, with 74.94%, 75.1%, and 74.58% for P-M, R-M, and F1-M, respectively; the BE-CRE model is obtained after fusing entity representations on the basis of BERT-Softmax, and its extraction results are further improved, with P-M, R-M, and F1-M improved by 3.61%, 4.55%, and 3.73%, respectively.

Entity representation was superior to conventional attention mechanisms in this context because it explicitly captures the unique semantic properties of entities, enabling the model to better understand complex domain-specific relationships, which attention mechanisms alone may overlook.In summary, the BE-CRE model performed best on the public CharacterCRE dataset with the highest P-M, R-M, and F1-M values of 78.55%, 79.65%, and 78.31%, respectively, which proved the generalization and validity of BE-CRE on different domain datasets.

Table 6. Extraction Results of All Models on Character CRE Dataset

Models	P-M/%	R-M/%	F1-M/%
BiLSTM-ATT	55.26	54.25	54.7
BiGRU-ATT	54.24	50.64	51.38
BiLSTM-2ATT	59.72	59.9	59.81
BiGRU-2ATT	57.78	55	56.36
ALBERT-softmax	69.43	72.42	70.08
RoBERTa-softmax	71.22	71.72	70.82
BERT-softmax	74.94	75.1	74.58
BE-CRE	78.55	79.65	78.31

Note: P-M: Precision-Macro (%); R-M: Recall-Macro (%); F1-M: F1-Macro (%). ATT: Attention mechanism; 2ATT: Dual-attention mechanism.

DISCUSSION

The effectiveness of relation extraction methods depends on the ability to capture contextual semantic information in text features (Zhao *et al*., 2024). Our experiments show that the BE-CRE model outperforms traditional attention-based models like BiLSTM-ATT and BiGRU-ATT, achieving substantial improvements in F1-M scores from 67.28% and 65.23% to 89.29%, representing improvements of 22.01 and 24.06 percentage points respectively. This significant improvement is due to BE-CRE's integration of BERT's contextual representations with explicit entity information (Devlin, *et al*., 2020; Shi and Lin, 2019). In contrast, single-layer attention mechanisms in BiLSTM-ATT and BiGRU-ATT fail to capture complex semantic

relationships in domain-specific texts, leading to F1-M scores below 70%.

The introduction of dual-attention mechanisms in BiLSTM-2ATT and BiGRU-2ATT showed notable improvements, with F1-M scores reaching 81.49% and 78.94% respectively (Chen *et al*., 2025; Dengfeng *et al*., 2025). This improvement demonstrates the value of incorporating both word-level and sentence-level attention for capturing hierarchical semantic features. However, these models still fall short of the BE-CRE model's performance, which achieved an F1-M score of 89.29% on the Wheat CRE dataset. This superior performance can be attributed to the model's unique integration of entity representations with BERT's contextual embeddings (Peters *et al*., 2019).

The construction of the Wheat CRE dataset through remote supervision with manual correction proved to be an effective approach for addressing the data scarcity challenge (Guo *et al*., 2022). The dataset's quality is reflected in the consistent performance across different relation categories, with most categories achieving F1-M scores above 85%. However, the uneven distribution of relation categories presents challenges for model training. As shown in Table 1, the "Distribution Range" category dominates with 523 samples (31.1% of total), while "Genus Orders" has only 199 samples (11.8%). This imbalance contributes to performance variations, with the relatively lower performance (79.25%) for the "Damage Crops" relationship category highlighting the ongoing challenge of handling complex sentence structures and nested entities, suggesting areas for future improvement in both dataset construction and model architecture (Wang *et al*., 2025).

The comparative analysis of different pre-trained models (ALBERT, RoBERTa, and BERT) revealed a trade-off between performance and computational efficiency (Lan *et al*., 2019; Liu *et al*., 2019). While ALBERT and RoBERTa performed competitively with F1-M scores above 84%, the base BERT model outperformed them with an F1-M of 87.99% before integrating entity representations. The BE-CRE model further improved to 89.29% F1-M, demonstrating the effectiveness of combining entity representations with contextual embeddings, with a computational cost increase of only 1.7 seconds per epoch (from 21.75s to 23.45s).

The model's generalization capability is particularly noteworthy, as demonstrated by its performance on the Character CRE dataset (Han *et al*., 2020). Achieving the highest F1-M score of 78.31% on this general-domain dataset suggests that the principles underlying the BE-CRE model's architecture are not limited to agricultural applications but can be effectively applied to broader relation extraction tasks. This generalization ability is particularly important for developing robust NLP systems that can handle both domain-specific and general-purpose tasks (Hupkes *et al*., 2023).

Despite these promising results, several limitations and opportunities for future research emerge from our study. The uneven distribution of relation categories in the WheatCRE dataset, with the largest category ("Distribution Range") containing 2.6 times more samples than the smallest ("Genus Orders"), remains a challenge for model training despite our stratified sampling strategy. Additionally, the current set of six relation categories, while covering major aspects of wheat diseases and pests' relationships, could be expanded to capture more nuanced relationships within the domain, such as temporal relationships, causal relationships, and treatment-pathogen interactions.

Future research directions could explore several promising avenues. First, the integration of domain-specific knowledge bases could further enhance the model's ability to capture specialized semantic relationships (Li, D. *et al.*, 2020). Second, the development of more sophisticated entity representation methods could potentially improve the model's handling of complex nested entities and long-range dependencies, particularly for challenging categories like "Damage Crops" which showed lower performance due to complex sentence structures. Finally, the application of advanced data augmentation techniques could help address the challenge of uneven category distribution while maintaining data quality, potentially improving performance on underrepresented categories.

In conclusion, our study highlights the effectiveness of combining BERT-based contextual representations with explicit entity information for domain-specific relation extraction. The BE-CRE model's superior performance on both domain-specific and general datasets validates our approach to addressing the unique challenges of agricultural text processing. The model demonstrates significant improvements over existing approaches, with F1-M improvements ranging from 7.8 to 24.06 percentage points compared to baseline models. Although there are areas for improvement, especially in handling complex sentence structures and expanding relation categories, the current results lay a strong foundation for future advancements in agricultural knowledge graph construction and information extraction systems.

Conclusions: This study developed Wheat CRE dataset (1,681 samples across six relationship categories) and BE-CRE model for Chinese entity relation extraction in wheat diseases and pests. BE-CRE achieved F1-Macro scores of 89.29% on Wheat CRE and 78.31% on Character CRE, significantly outperforming baseline models including BiLSTM-Attention (67.28%), BiGRU-Attention (65.23%), and BERT-Softmax (87.99%). The results demonstrate that incorporating entity-specific

information enhances relation extraction accuracy compared to attention-only mechanisms. This research provides foundational infrastructure for intelligent agricultural systems, supporting automated disease identification and precision agriculture applications.

Authors' contribution: All authors interpreted the data, critically revised the manuscript for important intellectual contents and approved the final version.

REFERENCES

Chen, J., T. Zhang, Z. Yan, Z. Zheng, W. Zhang and J. Zhang (2025). Attention-based BiLSTM with positional embeddings for fake review detection. J. Big Data 12(1): 83. https://doi.org/10.1186/s40537-025-01130-9

Chen, Z., X. Chen, Z. Yang, J. He, H. Bai and Y. Liu (2024). A Weighted Diffusion Graph Convolutional Network for Relation Extraction. J. Electr. Comput. Eng. 2024: 8729621. https://doi.org/10.1155/2024/8729621

Dengfeng, Z., T. Chaoyang, F. Zhijun, Z. Yudong, H. Junjian and H. Wenbin (2025). Multi scale convolutional neural network combining BiLSTM and attention mechanism for bearing fault diagnosis under multiple working conditions. Sci. Rep. 15(1): 13035. https://doi.org/10.1038/s41598-025-96137-w

Dolatabadian, A. *et al*. (2025). Image-based crop disease detection using machine learning. Plant Pathol. 74(1): 18-38. https://doi.org/10.1111/ppa.14006

Devlin, J., M.W. Chang, K. Lee and K. Toutanova (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proc. NAACL-HLT 2019, pp. 4171-4186. https://doi.org/10.18653/v1/N19-1423

Guo, X., X. Zhang, Y. Geng, J. Zhang, M. Zhang and Y. Zhang (2022). CG-ANER: Enhanced Contextual Embeddings and Glyph Features-Based Agricultural Named Entity Recognition. Comput. Electron. Agric. 194: 106776. https://doi.org/10.1016/j.compag.2022.106776

Han, X., Y. Zhang, W. Zhang and T. Huang (2020). An attention-based model using character composition of entities in Chinese relation extraction. Information 11(2): 79. https://doi.org/10.3390/info11020079

Han, Z. and J. Wang (2024). Knowledge enhanced graph inference network based entity-relation extraction and knowledge graph construction for industrial domain. Front. Eng. Manag. 11(1): 143-158. https://doi.org/10.1007/s42524-023-0273-1

Hossain, S.M.M. and K. Deb (2021). Proc. ICO 2020, AISC 1324, pp. 530-545. Plant Leaf Disease Recognition Using Histogram-Based Gradient Boosting Classifier. https://doi.org/10.1007/978-3-030-68154-8_47

Hupkes, D., M. Giulianelli, V. Dankers, M. Artetxe, Y. Elazar, T. Pimentel, C. Christodoulopoulos, A. Rozen, T. Linzen, D. Baroni and Z. Belinkov (2023). A taxonomy and review of generalization research in NLP. Nat. Mach. Intell. 5(10): 1161-1174. https://doi.org/10.1038/s42256-023-00729-y

Hussain, D. *et al*. (2024). A review on identification characters and IPM of wheat aphid by using non-conventional methods. Int. J. Trop. Insect Sci. 44(2): 419-432. https://doi.org/10.1007/s42690-024-01162-3

Jalli, M. *et al*. (2021). Effects of crop rotation on spring wheat yield and pest occurrence in different tillage systems: a multi-year experiment in Finnish growing conditions. Front. Sustain. Food Syst. 5: 647335. https://doi.org/10.3389/fsufs.2021.647335

Joshi, P., K.S. Sandhu, G.S. Dhillon, J. Chen and K. Bohara (2024). Detection and monitoring wheat diseases using unmanned aerial vehicles (UAVs). Comput. Electron. Agric. 224: 109158. https://doi.org/10.1016/j.compag.2024.109158

Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma and R. Soricut (2019). Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942. https://doi.org/10.48550/arXiv.1909.11942

Li, D., Y. Zhang and D. Li (2020). Review of Entity Relation Extraction Methods. J. Comput. Res. Dev. 57(7): 1424-1448. https://doi.org/10.7544/issn1000-1239.2020.20190358

Li, Q., W. Li, J. Zhang, Y. Zheng, J. Chen, Y. Yu and X. Su (2024). Joint extraction of entity and relation based on fine-tuning BERT for long biomedical literatures. Bioinform. Adv. 4(1): vbae194. https://doi.org/10.1093/bioadv/vbae194

Li, Z., Y. Lian, X. Ma, X. Zhang and C. Li (2020). Bio-semantic relation extraction with attention-based external knowledge reinforcement. BMC Bioinformatics 21(1): 213. https://doi.org/10.1186/s12859-020-3540-8

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen and V. Stoyanov (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692

Long, M., M. Hartley, R.J. Morris and J.K. Brown (2023). Classification of wheat diseases using deep learning networks with field and glasshouse

images. Plant Pathol. 72(3): 536-547. https://doi.org/10.1111/ppa.13684

Lv, X., X. Li, P. Li, D. Tao and L. Shen (2022). Chinese Named Entity Recognition in the Geoscience Domain Based on BERT. Earth Space Sci. 9(2): e2021EA002166. https://doi.org/10.1029/2021EA002166

Nanavaty, A. *et al*. (2024). Integrating deep learning for visual question answering in Agricultural Disease Diagnostics: Case Study of Wheat Rust. Sci. Rep. 14(1): 28203. https://doi.org/10.1038/s41598-024-79793-2

Pacal, I. *et al*. (2024). A systematic review of deep learning techniques for plant diseases. Artif. Intell. Rev. 57(11): 304. https://doi.org/10.1007/s10462-024-10944-7

Peters, M.E., M. Neumann, R.L. Logan IV, R. Schwartz, V. Joshi, S. Singh and N.A. Smith (2019). Knowledge enhanced contextual word representations. arXiv preprint arXiv:1909.04164. https://doi.org/10.48550/arXiv.1909.04164

Shi, P. and J. Lin (2019). Simple BERT Models for Relation Extraction and Semantic Role Labeling. arXiv preprint arXiv:1904.05255. https://doi.org/10.48550/arXiv.1904.05255

Shim, D.S. and J. Shim (2023). A modified stochastic gradient descent optimization algorithm with random learning rate for machine learning and deep learning. Int. J. Control Autom. Syst. 21(11): 3825-3831. https://doi.org/10.1007/s12555-022-0947-1

Wang, H. and R. Zhao (2024). Knowledge graph of agricultural engineering technology based on large language model. Displays 85: 102820. https://doi.org/10.1016/j.displa.2024.102820

Wang, L., F. Wu, X. Liu, J. Cao, M. Ma and Z. Qu (2025). Relationship extraction between entities with long distance dependencies and noise based on semantic and syntactic features. Sci. Rep. 15(1): 15750. https://doi.org/10.1038/s41598-025-00915-5

Wu, S.S. *et al*. (2020). Construction of visualization domain-specific knowledge graph of crop diseases and pests based on deep learning. Trans. Chin. Soc. Agric. Eng. 36(24): 177-185. https://doi.org/10.11975/j.issn.1002-6819.2020.24.021

Xu, L. *et al*. (2023). Wheat leaf disease identification based on deep learning algorithms. Physiol. Mol. Plant Pathol. 123: 101940. https://doi.org/10.1016/j.pmpp.2022.101940

Yao, X., X. Hao, R. Liu, L. Li and X. Guo (2024). AgCNER, the First Large-Scale Chinese Named Entity Recognition Dataset for Agricultural Diseases and Pests. Sci. Data 11: 769. https://doi.org/10.1038/s41597-024-03578-5

Yu, B. *et al*. (2020). Joint entity and relation extraction with set prediction networks. Proc. ECAI 2020, pp. 2282-2289, Amsterdam, Netherlands. https://doi.org/10.3233/FAIA200356

Yue, Y. *et al*. (2020). Agricultural pest and disease relation extraction based on multi-attention mechanism and distant supervision. J. Anhui Agric. Univ. 47(4): 682-686. https://doi.org/10.13610/j.cnki.1672-352x.20200907.020

Zhang, H., H. Si, X. Ma, L. Xi and X. Xu (2021). Research and Application of Agriculture Knowledge Graph. Proc. 5th Int. Conf. Electronic Information Technology and Computer Engineering (EITCE 2021), pp. 680-688, Xiamen, China. https://doi.org/10.1145/3501409.3501531

Zhao, X., Y. Deng, M. Yang, L. Wang, R. Zhang, H. Cheng, W. Lam, Y. Shen and R. Xu (2024). A Comprehensive Survey on Relation Extraction: Recent Advances and New Frontiers. ACM Comput. Surv. 57(4): Article 89. https://doi.org/10.1145/3674501

Zhao, Y., X. Yuan, Y. Yuan, S. Deng and J. Quan (2023). Relation extraction: advancements through deep learning and entity-related features. Soc. Netw. Anal. Min. 13(1): 92. https://doi.org/10.1007/s13278-023-01095-8.