agronomy

MDPI

Article

Knowledge Graph Construction and Representation Method for Potato Diseases and Pests

Wanxia Yang, Sen Yang, Guanping Wang *, Yan Liu, Jing Lu and Weiwei Yuan

Mechanical and Electrical Engineering College, Gansu Agricultural University, Lanzhou 730070, China; yangwanxia@gsau.edu.cn (W.Y.); yangsen@gsau.edu.cn (S.Y.); liuyan@gsau.edu.cn (Y.L.); luj@st.gsau.edu.cn (J.L.); yuanww@st.gsau.edu.cn (W.Y.)
* Correspondence: wangguanping@gsau.edu.cn

Abstract: Potato diseases and pests have a serious impact on the quality and yield of potatoes, and timely prevention and control of potato diseases and pests is essential. A rich knowledge reserve of potato diseases and pests is one of the most important prevention and control measures; however, valuable knowledge is buried in the massive data of potato diseases and pests, making it difficult for potato growers and managers to obtain and use it in a timely manner and to develop the potential of knowledge. Therefore, this paper explores the construction method of a knowledge graph for automatic knowledge extraction, which extracts the knowledge of potato diseases and pests scattered in heterogeneous data from multiple sources, organises it into a semantically related knowledge base, and provides potato growers with professional knowledge and timely guidance to effectively prevent and control potato diseases and pests. In this paper, a data corpus on potato diseases and pests, called PotatoRE, is first constructed. Then, a model of ALBert-BiLSTM-Self_Att-CRF is designed to extract knowledge from the corpus to form a triplet structure, which is imported into the Neo4j graph database for storage and visualisation. Furthermore, the performance of the model constructed in this paper is compared and verified using the datasets PotatoRE and People's Daily. The results show that compared to the SOTA models of ALBert BiLSTM-CRF and ALBert BiGRU-CRF, the accuracy of our model has been improved by 2.92% and 3.12%, respectively, using PotatoRE. Compared to the Bert BiLSTM-CRF model on two datasets, our model not only improves the accuracy, recall, and F1 values, but also has a higher efficiency. The model in this paper solves the problem of the difficult recognition of nested entities. On this basis, through comparative experiments, the TransH model is used to effectively represent the constructed knowledge graph, which lays the foundation for achieving inference, extension, and automatic updating of the knowledge base. The achievements of the thesis have made certain contributions to the automatic construction of large-scale knowledge bases.

Keywords: potato diseases and pests; knowledge graph; knowledge extraction; knowledge representation; named entity recognition

check for updates

Citation: Yang, W.; Yang, S.; Wang, G.; Liu, Y.; Lu, J.; Yuan, W. Knowledge Graph Construction and Representation Method for Potato Diseases and Pests. *Agronomy* **2024**, *14*, 90. https://doi.org/10.3390/agronomy14010090

Academic Editor: Francis Drummond

Received: 26 November 2023
Revised: 27 December 2023
Accepted: 28 December 2023
Published: 29 December 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Potato diseases and pests are known to be one of the most important factors affecting the quality and yield of potatoes. The Great Famine in Ireland caused by potato blight in the mid-19th century is the best example [1]. Therefore, the prevention and control of potato diseases and pests is essential. However, there is a lack of specialised, systematic knowledge among potato growers, and authoritative knowledge on potato diseases and pest control is mostly contained in books or on certain agricultural websites, making such knowledge not only dispersed, but also not correlated, thus making it inaccessible to potato growers and managers and less conducive to systematic application. Therefore, it is necessary to study the construction of knowledge graphs. Because the construction of a knowledge graph can extract knowledge from fragmented, unrelated, and multi-source heterogeneous data and organise knowledge in the form of "entity-relationship-entity"

triplets and corresponding "attribute-value-pairs", it provides an important way to quickly query and obtain systematic expert knowledge [2]. For example, when a grower finds that potato leaves show brown spots with irregular spot texture, no concentric whorls, and a white mould on the spots, he or she inputs these features into the potato pests and diseases knowledge base for querying. Based on the semantic association, the corresponding disease name, i.e., potato late blight, as well as the cause of the disease, its relationship with the environment and the corresponding control methods will be obtained. This will help potato growers to correctly understand the causes of the disease and treat it as soon as possible. However, due to the complex structure and variety of potato pest and disease data, it brings great challenges to the automatic construction of a potato pests and diseases knowledge map. And there are few research reports on the construction and application of a potato pests and diseases control knowledge map. Therefore, this article aims to construct a knowledge graph of potato pests and diseases, and studies entity naming recognition and relationship extraction methods based on deep learning technology. A model is established to extract scattered potato pests and diseases knowledge from a large amount of data, forming a knowledge base with semantic relationships. Therefore, with the goal of constructing a knowledge map of potato pests and diseases, this paper researches the entity naming recognition and relationship extraction method based on deep learning technology and establishes a model to extract the fragmented knowledge of potato pests and diseases from data to form a knowledge base with semantic relationships. It can provide auxiliary decision-making assistance for agricultural experts, provide knowledge guidance for potato growers or managers, and also provide knowledge support for the intelligent question answering and recommendation system of the agricultural intelligent service platform [3,4], so as to make the static knowledge flow, apply it to agricultural production practice quickly, and give great play to the value of the knowledge, which is of good theoretical and application value.

Name entity recognition (NER) and relationship extraction (RE) are two core tasks in knowledge graph construction [5,6]. With the development of deep learning technologies, NER and RE models based on recurrent neural networks [7,8] and bidirectional encoder representation of transformers (BERT) [9] have performed well and become the main method for knowledge extraction. They are widely used in the automatic construction of knowledge bases in agriculture, medicine, finance, and other fields. For example, the NER model developed in [10], which is based on RNN networks, can accurately recognise entities in disease and pest data. The researchers found that extracting features by combining word embeddings with neural networks resulted in better recognition of entities and relationships. Zhang et al. [11] constructed an entity recognition model using word2vec and BiLSTM-CRF to significantly improve the performance of the original model. However, word2vec is static and cannot solve the problem of polysemy. For this purpose, BERT word embedding [12,13] was introduced. Reference [14] combines the character-level features of BERT representations with external dictionary features for NER in the agricultural field with a recognition accuracy of 94.84%. However, the BERT model generates a large number of parameters during the training process and results in time-consuming training. For this reason, the Reduced Bert (ALBert) model [15] has been developed, which mainly saves model parameter space by sharing parameters. For example, Chen et al. used ALBert not only to improve the effectiveness of NER, but also to ensure efficiency. The BiLSTM model can capture global contextual information, but it cannot capture long-range related information. Therefore, the attention mechanism is used to solve this problem. For example, the introduction of an attention mechanism in BiLSTM-CRF has been successfully applied to entity extraction tasks in the agricultural domain [16]. Guo et al. constructed the Att-BiLSTM-CRF model for pest and disease feature extraction, which solved the problem of missing inner semantic information [17].

In summary, by introducing the efficient ALBert model and attention mechanism into the BiLSTM-CRF model, a multilayer network knowledge joint extraction model is constructed to extract complex nested knowledge from the potato pest and disease corpus in

the paper. This method effectively balances the efficiency and complexity of the model, provides a scheme for the automatic construction of large-scale knowledge graphs, and also provides experimental guidance for the construction of multilayer models.

Furthermore, compared to knowledge graph construction, knowledge representation learning [18] is more important. It can not only explore the implicit relationship between disease and pest entities, so as to infer new knowledge, but also discover the interactions that exist between different pests and diseases to better predict and prevent the occurrence of diseases and pests. Therefore, knowledge representation learning, based on the constructed knowledge graph, is further explored to achieve knowledge inference and fusion. In turn, automatic knowledge updating can be achieved, which greatly enhances the value of the knowledge graph.

2. Materials and Methods

2.1. Process of Construction and Representation of the Knowledge Graph in This Paper

The key elements in the construction of a knowledge graph are the quantity and quality of data. Therefore, a high-quality corpus of potato disease and pest data (PotatoRE) is first constructed in this paper. A deep learning model is then developed to extract entities and relationships from the unstructured data in this corpus to obtain triplet knowledge, which is stored in the Neo4j graph database to form a knowledge graph. Further, the performance of knowledge representation models TransE [19], TransH [20], TransR [21], and TransD [22] are compared in the experiment, and the better one is selected to further represent and optimise the knowledge graph of potato diseases and pests. The main process of knowledge graph construction and representation in this paper is shown in Figure 1.

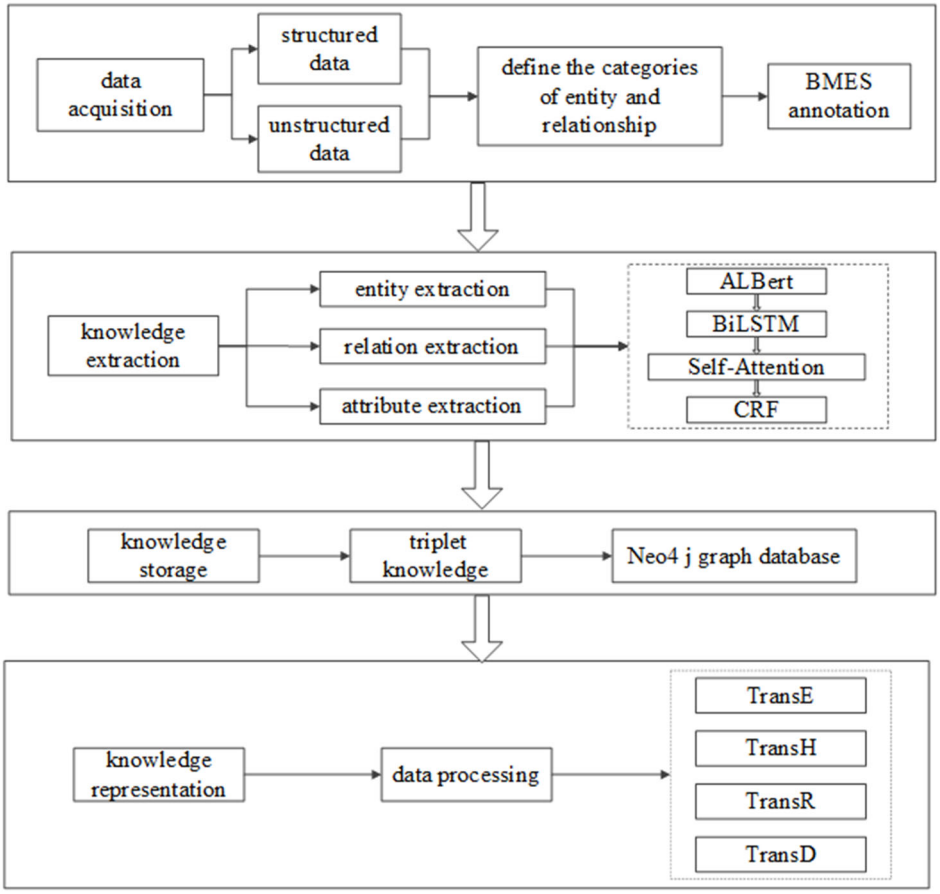


Figure 1. Process of construction knowledge graph of potato diseases and pests.

2.2. Design of Entity and Relationship Joint Extraction Model

It can be seen in Figure 1 that entity and relationship extraction is the most important link in the construction of a knowledge graph. Therefore, a layered model based on ALBert word embedding and the BiLSTM network was designed to extract the entity and relationship of potato diseases and pests from the corpus, as shown in Figure 2.

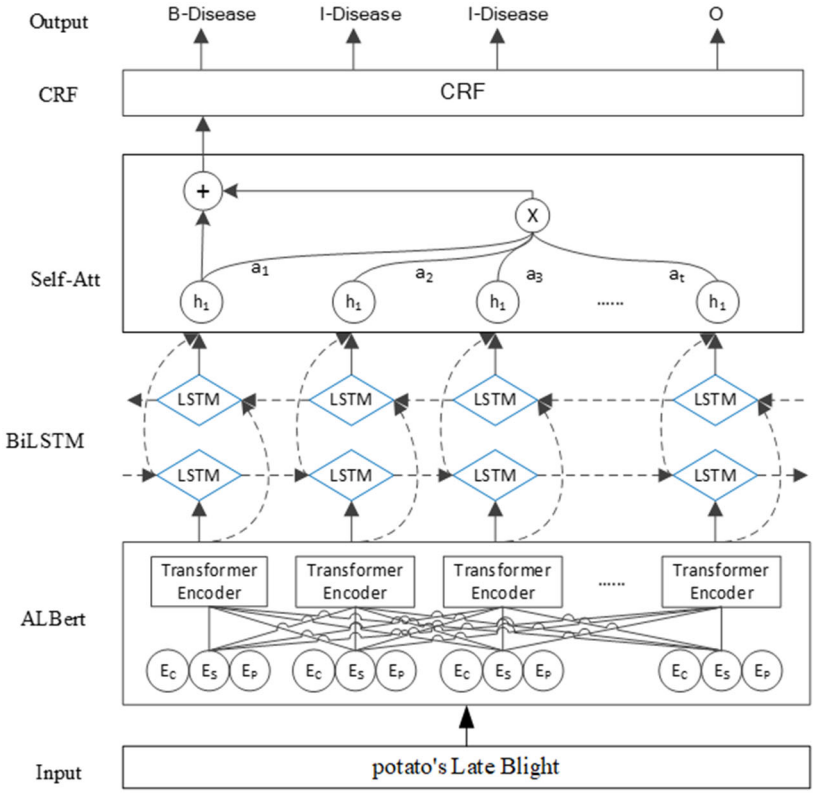


Figure 2. Knowledge extraction model structure.

In Figure 2, ALBert converts the input text into a word vector, and its structure is still a transformer. However, it has made innovations in factorisation, parameter embedding, cross-layer parameter sharing, and inter-sentence coherence loss, which reduce the number of parameters and improve the stability of the model. In particular, ALBert generates character vectors based on contextual information and effectively solves the problem of polysemy. The BiLSTM layer of the model obtains information about the previous and next moments by means of forward and backward propagation parameters. The self-attention mechanism of the model captures the internal semantic relationships of sentences and improves the accuracy of knowledge extraction. The Conditional Random Field (CRF) [23] serves as a decoder to output the order of labels based on sequence characteristics. The label sequence with higher scores is obtained using the Viterbi algorithm, and SoftMax normalisation is further performed on the label as the final output result.

2.3. Knowledge Representation Algorithms

The translation model TransX is chosen to vectorise the knowledge graph of potato diseases and pests in this paper, and the low-dimensional dense vectors of entities and relationships are obtained so that the constructed knowledge graph can be better applied to downstream tasks such as information retrieval, question-answering systems, and intelligent recommendations.

Knowledge graph representation based on translation is an independent modelling method based on triples [24]. The relationship (r) is considered a transformation from the head (h) to the tail (t) in each triplet (head, relationship, tail). By continuously adjusting h, r, and t, (h + r) is made as close as possible to t, i.e., $h + r \approx t$. The first proposed TransE

translation model (as shown in Figure 3a) is simple and efficient, but it cannot handle the relationships between 1 to N, N to 1, and N to N. For this purpose, TransH is proposed (as shown in Figure 3b), which allows an entity to have different distributed representations when different relationships are involved. However, TransE and TransH cannot overcome the limitations of representing entities and relationships in the same semantic space. So TransR (shown in Figure 3c) is introduced, which can model entities and relationships in different spaces to form entity and relationship spaces, where they can be transformed. However, TransH and TransR project entities are based only on relationships and ignore the diversity of entities. In addition, TransH and TransR use matrix-vector multiplication for projection, which increases the time complexity. Therefore, TransD proposes a dynamic mapping matrix based on entities and relationships, which can take into account the diversity of entities and relationships (as shown in Figure 3d). To better represent the knowledge graph constructed in this paper, the effects of four representation methods on this knowledge graph are compared to find a better knowledge representation method.

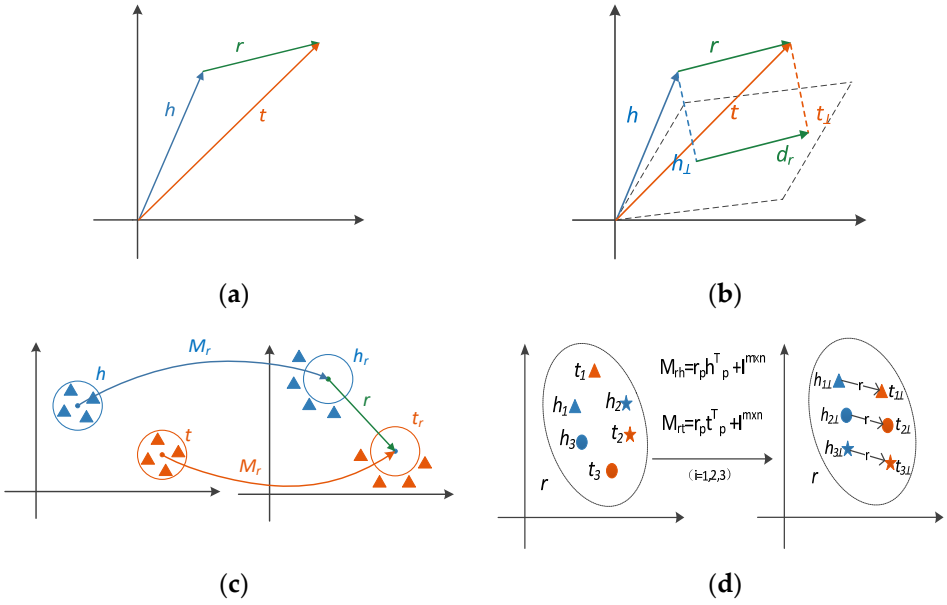


Figure 3. Knowledge representation methods. (**a**) TransE; (**b**) TransH; (**c**) TransR; (**d**) TransD.

3. Discussion

3.1. Experimental Environment

An Intel (R) Core (TM) i7-8750H CPU @ 2.20 GHz, GPU NVIDIA GTX 1050Ti hardware device with a disk size of 2 TB was selected in this experiment. The development language was Python 3.7, the compiler was Pycharm v.2022, and the deep learning development platform was Tensorflow v.1.3. Its efficient, fast, simple, and convenient characteristics greatly improved the training speed of the model.

3.2. Construction of Potato Disease and Pest Data Corpus

The data of the PotatoRE constructed in this paper were mainly obtained from the Internet and agricultural textbooks, such as the National Agricultural Science Data Center (https://www.agridata.cn/, accessed on 5 September 2022), the China Agricultural Information Network (http://www.agri.cn/, accessed on 5 September 2022), the Agricultural Encyclopedia and Wikipedia, as well as digital textbooks such as *Prevention and Control of Potato Diseases and Pests*, *Chinese Crop Diseases and Pests*, and *Identification and Control of Potato Diseases and Pests*. A small amount of structured data from the National Agricultural Science Data Centre could be directly converted into triplets. The data from various websites, such as the Encyclopedia, was mainly semi-structured data that was crawled using scraping. The data in agricultural books was unstructured and made up the major-

ity. Semi-structured and unstructured data require pre-processing such as cleaning, noise reduction, and redundancy removal before knowledge can be extracted to provide high-quality data for annotation. Data annotation is a necessary step in using deep learning methods for knowledge extraction. It mainly involves defining appropriate labels based on entity types, and then using appropriate annotation methods to annotate the entities in unstructured data. Based on the definitions and concepts in the two professional books *Chinese Agri-cultural Thesaurus* and *Special Classification for Agriculture* and under the guidance of Mr. Zhang Tingwei, a plant protection expert, the entities in the dataset of this paper were classified into six categories: diseases, prevention and control, symptoms, etiology, onset locations, and distribution area. After annotation and manual calibration, a potato pest and disease corpus of 1500 sentences and 202,000 characters was obtained for subsequent experiments.

Next, the BMES annotation method with better performance by comparative experiments was selected to annotate the corpus dataset using the YEDDA tool [25]. The example of annotation is shown in Figure 4b. After organising the annotated dataset, a dataset consisting of 6 entity types, 8 relationships, and 4 attribute types, with a total of 8971 entities, 5238 relationships, and 3464 attribute samples, was obtained for the subsequent experiments. Detailed information about the experimental dataset is given in Table 1.

Table 1. The experimental dataset.

Entity Type	Quantity	Relationship Type	Quantity	Attribute Type	Quantity
Disease	579	Pest damage	540	Onset cycle	579
Prevention and control method	1153	Pest control	500	Prevention and control	981
Symptom	1472	Disease control	1012	Onset condition	876
Etiology	2452	Disease damage	987	Toute of transmission	1028
Onset location	2966	Another name	1042		
Distribution area	200	Distribution area	463		
		Disease department	333		
		Pest department	324		
Total number of entities	8971	Total number of relationships	5238	Total number of attributes	3464

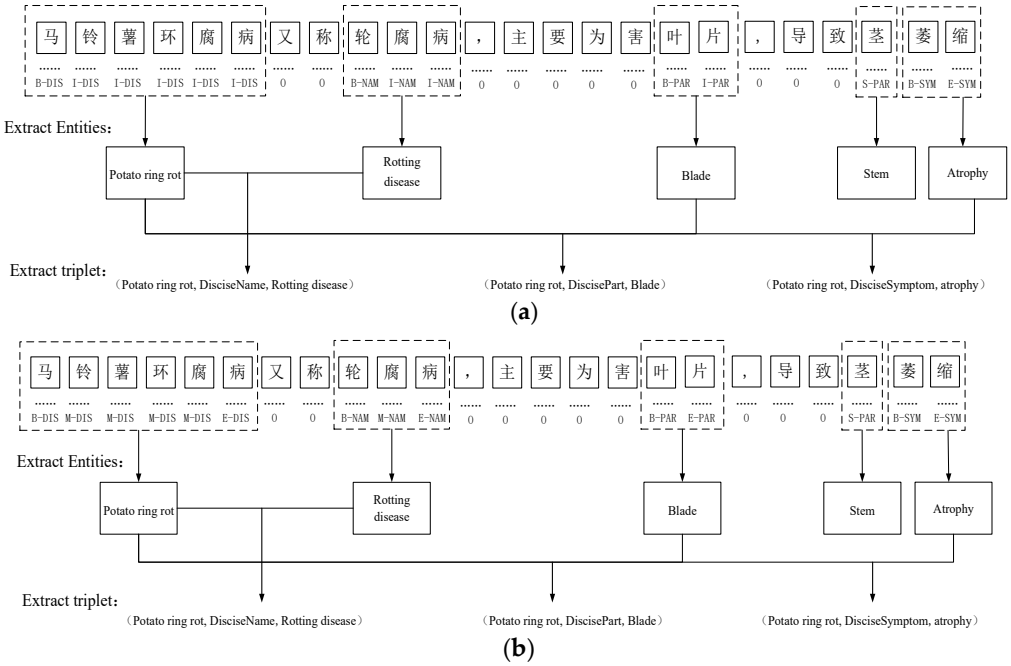
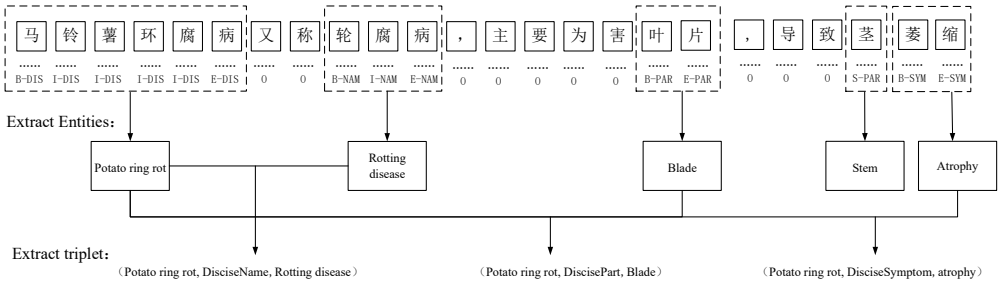


Figure 4. *Cont.*



(**c**)

Figure 4. Example of annotation. DIS, NAM, PAR, and SYM are the abbreviations of the labels: (**a**) Example of BIO Annotation; (**b**) Example of BMES Annotation; (**c**) Example of BIOES Annotation. (DIS—Discise Name; NAM—Another Name; PAR—Discisepart; SYM—Discise Symptom; The definition of "马铃薯环腐病又称轮腐病, 主要为害叶片, 导致茎萎缩" is "Potato ring rot, also known as rotting disease, mainly affects the leaves and causes stem atrophy").

3.3. Comparative Experiment and Analysis of Sequence Annotation Method

Before a model can be used to extract knowledge from a dataset, sequence annotation is required. The correct selection of annotation methods and the consistency of the annotation results directly determine the accuracy of entity recognition and relationship extraction. Therefore, this experiment investigated the performance of three sequence annotation methods, such as BIO, BMES, and BIOES, on the dataset, and the BMES annotation method was found to be the best and was selected for knowledge extraction. The comparative experimental model was set as Word2vec-BiLSTM-CRF [26].

The specific meaning of the annotation method is as follows: In the BIO annotation method, B-disease represents the beginning of the entity disease, I-disease represents the middle or end of the entity disease, and O indicates non-entity (Figure 4a). In the BMES annotation, B, M, E, and S represent the beginning position, middle position, end position, and a single word, respectively (Figure 4b). In the BIOES annotation method, B, I, O, and E represent the beginning, middle, non-entity, and end of the entity, respectively, while S indicates that the word itself is an entity (Figure 4c). The annotation example is shown in Figure 4. In contrast, the BMES and BIOES methods can annotate entities more finely, which is superior to the BIO method.

Accuracy (*P*), Recall (*R*), and *F*1 values were selected as evaluation indicators for the experimental results. The specific calculations for these indices are shown in Formulas (1)–(3).

$$P = \frac{Number\ of\ potato\ disease\ and\ pest\ entities\ correctly\ identified}{Number\ of\ potato\ disease\ and\ pest\ entities\ identified} \times 100\%, \quad (1)$$

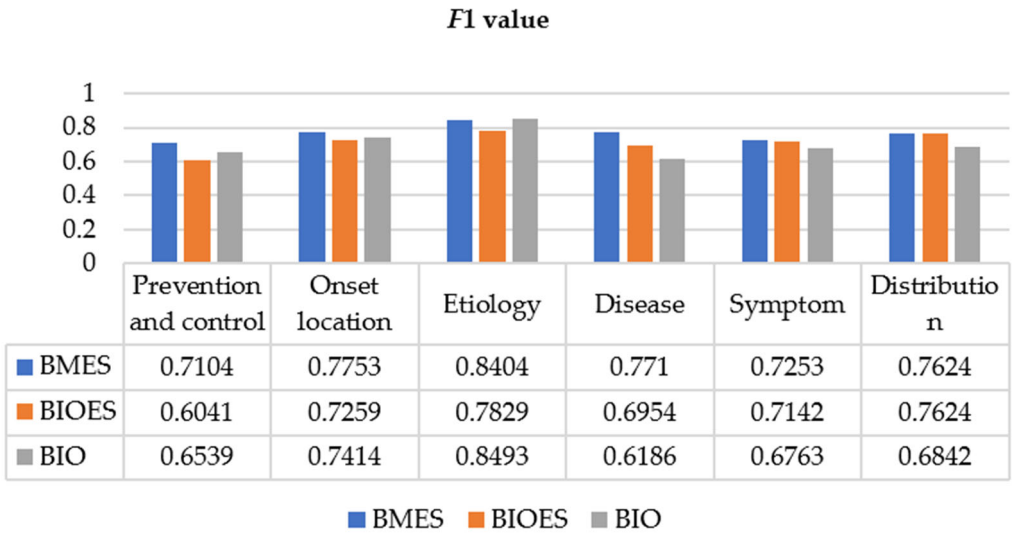
$$R = \frac{Number\ of\ potato\ disease\ and\ pest\ entities\ correctly\ identified}{Total\ number\ of\ potato\ pests\ and\ diseases\ entities} \times 100\%, \quad (2)$$

$$F1 = \frac{2 \times P \times R}{P + R}, \quad (3)$$

The experimental results of sequence annotation are shown in Table 2. The values in the Table 2 are the average of the recognition results for deferent types of entities such as prevention and control, onset location, etc. In addition, the *F*1 value is displayed graphically, as shown in Figure 5.

Table 2. Experimental results on the impact of annotation methods on entity recognition.

Annotation Method	*P*	*R*	*F*1
BIO	78.21	74.23	75.62
BMES	78.17	77.64	77.91
BIOES	71.31	71.48	71.24



	Prevention and control	Onset location	Etiology	Disease	Symptom	Distribution
BMES	0.7104	0.7753	0.8404	0.771	0.7253	0.7624
BIOES	0.6041	0.7259	0.7829	0.6954	0.7142	0.7624
BIO	0.6539	0.7414	0.8493	0.6186	0.6763	0.6842

Figure 5. *F*1 values for identification of each type of entity using different annotation methods.

It can be seen in Table 2 and Figure 5 that the *F*1 value of the BMES method is the highest, while the effect of the BIOES method is less effective. The reason for this is that BIOES annotations cannot correctly recognise some nested entities, resulting in missing annotations. The entity samples with missing annotations are considered negative samples by the model, which will ultimately lead to an increase in negative samples and have a negative impact on recognition performance. Compared to the BIO annotation, the BMES annotation adds an entity tail label "E" and a single entity annotation label "S", which has a positive impact on the training effectiveness and improves the model's recognition efficiency. Therefore, BMES was selected as the annotation method for the data in this paper.

3.4. Experiment and Analysis of Knowledge Extraction Model Performance

3.4.1. Analysis of the Relationship between Sample Size and Model Performance

The total number of samples for this experiment was divided equally into 10 parts, with the samples stacked sequentially to form 10 sub-samples, with each sample in the experiment divided into training, validation, and test sets in an 8:1:1 ratio. An experiment was then run to test the effect of the sample size on the performance of the BiLSTM and BiGRU models. When sample 1 was selected with a batch_size value of 8 and an epoch of 20, the accuracy of the ALBert-BiLSTM-CRF model and the ALBert-BiGRU-CRF model for knowledge extraction was 1. When the epoch was increased to 40, the accuracy of the ALBert-BiGRU-CRF model did not reach 1. When sample 2 was selected with a batch_size value of 8, regardless of the value of the epoch, the accuracy of both models never reached 1. When the sample size was increased to sample 4 with a batch_size value of 16, regardless of the number of epochs, the accuracy of the model did not reach 1. It can be seen that the number of training epochs, the size of the dataset, and the batch_size value can all cause the model to underfit or overfit, thereby affecting the results of knowledge extraction. Furthermore, the results show that the BiGRU model outperforms the BiLSTM model when other parameters remain unchanged and the sample size is small. However, when the number of samples is large, the performance of the BiLSTM model is better than that of the BiGRU model. The performance of the models is measured with precision, as shown in Figure 6.

Figure 6 shows that as the number of samples increases, the recognition accuracy of both models improves. When the number of samples increases to sample 4 and sample 5, the recognition performance of the two models is close. When the amount of training data continues to increase from sample 5 to the entire corpus, the recognition accuracy of the ALBert-BiLSTM-CRF model is still higher than that of the ALBert-BiGRU-CRF model. In theory, GRU simplifies the internal structure of LSTM, but in experiments, it

is found that BiGRU occupies more memory at runtime than the BiLSTM, resulting in a relatively longer training time for the BiGRU model. Therefore, the extraction knowledge from the potato pest and disease corpus constructed in this paper was based on the ALBert-BiLSTM-CRF model.

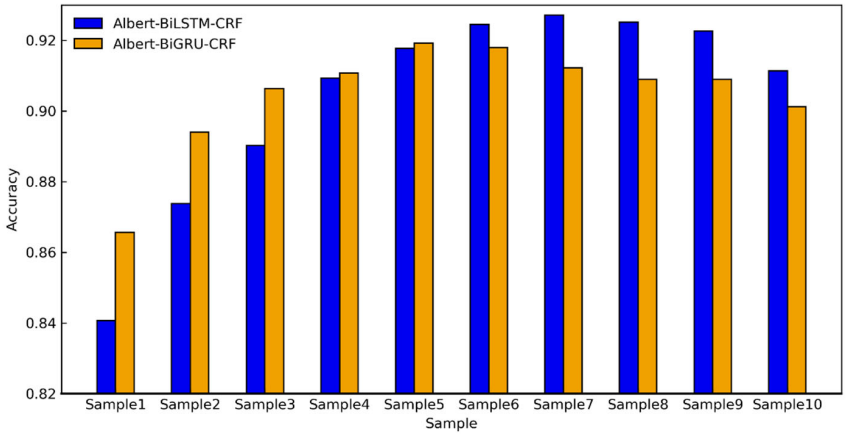


Figure 6. Influence of different sample sizes on model performance.

3.4.2. Comparison of Time Efficiency between ALBert and Bert Models

To test the time efficiency of the ALBert and Bert models, we compared the training time of the ALBert-BiLSTM-CRF, ALBert-BiGRU-CRF, Bert-BiLSTM-CRF, and Bert-BiGRU-CRF models. The epochs were set at 40 and the batch_size value was 4. The results are shown in Table 3.

Table 3. Comparison of training time.

Model	Training Time
ALBert-BiLSTM-CRF	8 h
Bert-BiLSTM-CRF	32 h
ALBert-BiGRU-CRF	10 h
Bert-BiGRU-CRF	36 h

Table 3 shows that for the same hyperparameter settings, the running time of the ALBert model is significantly shorter than that of the Bert model. At the same time, it also further confirms that the BiLSTM model is more efficient than the BiGRU model for this sample. Therefore, the ALBert and BiLSTM models were selected to build the full model.

3.4.3. Model Performance Comparison Experiment

To verify the robustness of the proposed ALBert-BiLSTM-Self_Att-CRF model, two sets of experiments were designed. The first set of experiments compared the performance of the proposed model with the other eight mainstream models using the dataset constructed in this paper (the results are shown in Table 4 and Figure 7). The second set of experiments compared the performance of the ALBert-BiLSTM-Self_Att-CRF model with the Global-Pointer and Word2vec-BiLSTM-CRF models using the public datasets People's Daily and Chinese Medical (the results are shown in Table 5). The epoch value of the two experimental groups was 40 and the batch_size value was 4. The ratio of dataset, validation set, and test set was 8:1:1.

Table 4 and Figure 7 show that the BiLSTM-CRF and BiGRU-CRF models significantly improved the recognition accuracy, recall, and F1 values on this corpus after adding the Bert module (Bert-BiGRU-CRF, Bert-BiLSTM-CRF). Moreover, their results are also better than Word2vec-BiLSTM-CRF. This is because the Bert layer can capture the semantics at the word and sentence level, and express the dynamic context of the sentences, thereby improving the generalisability of the model. Comparing the recognition results of the Bert-

CRF and Bert-BiLSTM-CRF models [27], the BiLSTM layer can fully extract features and identify entities more accurately. Moreover, comparing the experimental results of the Bert-BiLSTM-CRF and ALBert-BiLSTM-CRF, the entity recognition effect of ALBert on this corpus is better than that of the Bert model. This is because ALBert replaces the simpler NSP (Next Sentence Prediction) task with the SOP (Sentence Order Prediction) task. The results of the ALBert-BiLSTM-CRF model and the ALBert-BiLSTM-Self_Att-CRF model show that the knowledge extraction algorithm, after adding an attention mechanism, has improved the accuracy and *F*1 values by 2.92% and 3.22%, respectively, over the original model (without the attention mechanism). The reason is that the attention mechanism can capture the correlations within the sequence, which helps the model to identify rare entities based on the context. In particular, the result of the GlobalPointer model is also inferior to the model developed in this paper, although the GlobalPointer model has relatively good recognition of nested entities. Therefore, the ALBert-BiLSTM-Self_Att-CRF constructed in this paper has better recognition results on the PotatoRE corpus.

Table 4. Evaluation results of different models using the dataset constructed in this paper.

Model	Accuracy	Recall	F1
BiLSTM-CRF	0.7496	0.6697	0.7231
BiGRU-CRF	0.6654	0.6523	0.6542
Word2vec-BiLSTM-CRF	0.7417	0.7264	0.7791
Bert-CRF	0.7763	0.8169	0.7623
Bert-BiLSTM-CRF	0.7807	0.8042	0.7925
Bert-BiGRU-CRF	0.7821	0.7947	0.7887
ALBert-BiLSTM-CRF	0.8027	0.8081	0.7911
ALBert-BiGRU-CRF	0.8012	0.8137	0.8052
GlobalPointer	0.8164	0.8126	0.8081
ALBert-BiLSTM-Self_Att-CRF	0.8262	0.7879	0.8166

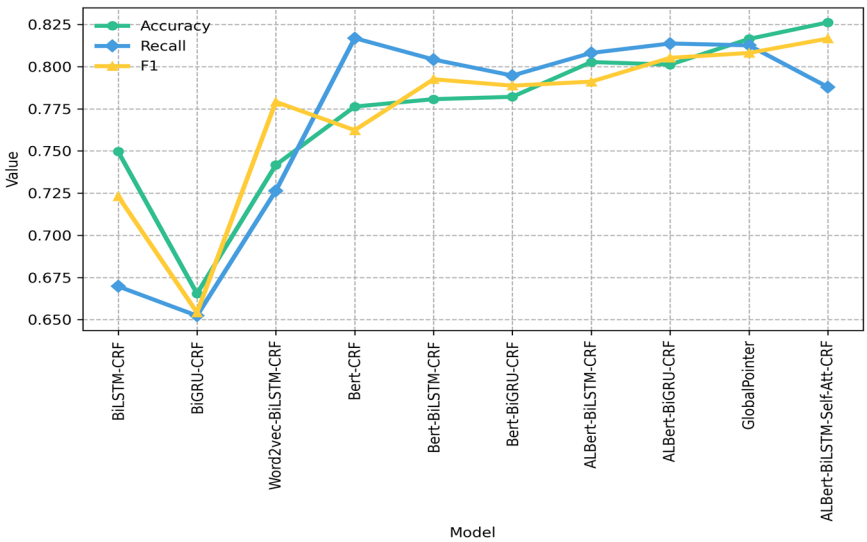
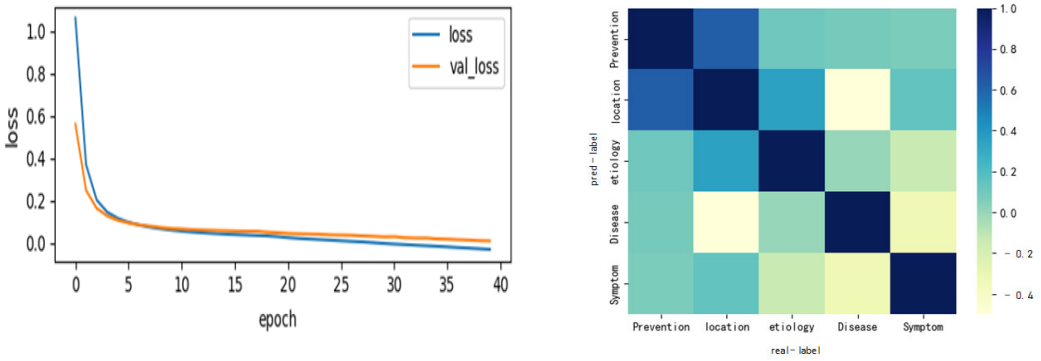


Figure 7. Evaluation results of different models.

Table 5. Evaluation results of different models using the public datasets.

	Dataset					
	People's Daily			Chinese Medical		
Model	**P**	**R**	**F1**	**P**	**R**	**F1**
Word2vec-BiLSTM-CRF	82.06	84.31	83.63	76.26	70.22	72.94
GlobalPointer	89.19	87.32	88.61	79.31	78.02	80.66
ALBert-BiLSTM-Self_Att-CRF	94.84	90.19	91.03	82.16	84.58	83.35

From Table 5, it can be seen that the model designed in this paper performs better than other mainstream models on the public dataset. In addition, as can be seen from the model loss curves in Figure 8a, and the correlation between predicted and true labels in Figure 8b, the ALBert-BiLSTM-Self_Att-CRF model designed has good generalisation capabilities on the training data.



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

Figure 8. (**a**) The loss function curve of ALBert-BiLSTM-Self_Att-CRF model; (**b**) the correlation between predicted and true labels.

3.4.4. Knowledge Storage

The triple structure knowledge formed using model extraction from the potato diseases and pests corpus was imported into the graph database for visualisation using Neo4j, as shown in Figure 9.

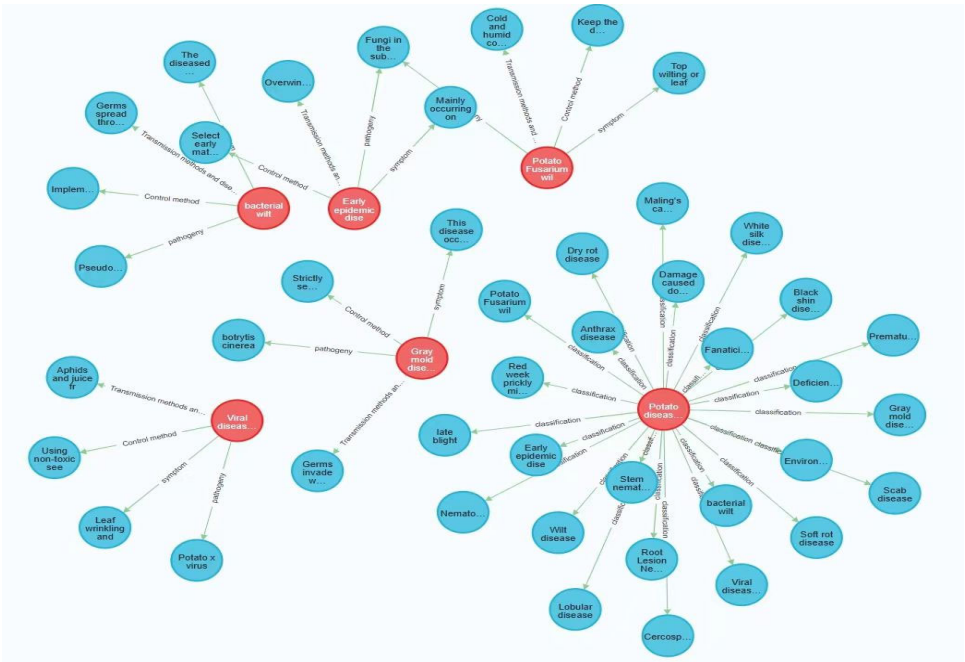


Figure 9. Visualisation of the potato disease and pest knowledge graph.

4. Results

In order to improve the representation and reasoning capabilities of the knowledge graph constructed in this paper, knowledge representation technology was further employed to enhance its scalability and increase its application value.

4.1. Experimental Dataset

APOC (A Package of Components) technology was used to export data from the Neo4j graph database as a CSV-type relationship file, which was then processed into entity txt files, relationship txt files, and triplet txt files. The entity file was a mapping of entities to entity IDs, with each line in the format of "entity\t entityID"; the relationship file was a mapping of relationships to relationship IDs, with each line in the format of "relationship\t relationshipID"; each line of the triplet file was in the format of "head entity\t tail entity\t relationship". Finally, the triplet file was divided into training, validation, and test sets in an 8:1:1 ratio.

4.2. Evaluation Metrics

In the knowledge graph representation learning algorithms, MRR (mean reciprocal rank) and Hit@N (such as Hit@10 and Hit@3) are typically used to evaluate the performance of the model. The MRR measures the average ranking score of correctly predicted entities, with a lower value indicating a higher ranking. Hit@N measures the probability of correct entities being ranked in the top N, with a higher value indicating better performance. Here, N was chosen as 10 and 3 to evaluate the model. The higher the MRR and Hit@N scores, the closer the model's predictions are to the actual values, and the better the model's performance.

4.3. Experimental Environment and Parameter Indicators

The experiment was run on the Windows 11 operating system using Python 3.7 and the PyCharm compiler, using the Tensorflow 1.14.0 learning framework. The model training parameters were that learning_rate was 0.001, the hidden_size (the length of the entity and relation word vectors) was 200, and the number of triplets entered in each epoch (batch_size) was 300. In the TransR model, we only used a hidden_sizeR of 10 and also kept the length of the relationship vector at 10 to ensure the model's performance.

4.4. Experiment and Results Analysis

The knowledge representation performances of the TransE, TransH, TransR, and TransD models on the constructed knowledge base were compared in the experiment, and the results are shown in Table 6 and their visualisation is shown in Figure 10. It can be seen that the Hit@10, Hit@3, and MRR metrics of the TransH model are all higher than the other three models, indicating that the TransH model is suitable for capturing and representing special semantic relationships in knowledge graphs. In addition, the TransH model can effectively deal with the different relationships between 1-to-N, N-to-1, and N-to-N situations involved in the entities in the knowledge graph, However, the metrics of the TransD model in the experimental results are low, which is somewhat different from the theoretical analysis of this model. A possible reason for this is that there were too few types of entity relationships in the experimental dataset, which is insufficient to fully demonstrate the performance of the model. Therefore, the TransH model was chosen to represent the knowledge graph of potato diseases and pests in this paper. During the training process, the loss rate of the TransH model with changes in epochs is illustrated in Figure 11, which shows that its loss rate can converge well.

Table 6. Comparison of knowledge representation models.

Model	Hit@10	Hit@3	MRR
TransE	0.503	0.492	0.416
TransH	0.524	0.516	0.437
TransR	0.47	0.457	0.389
TransD	0.461	0.448	0.366

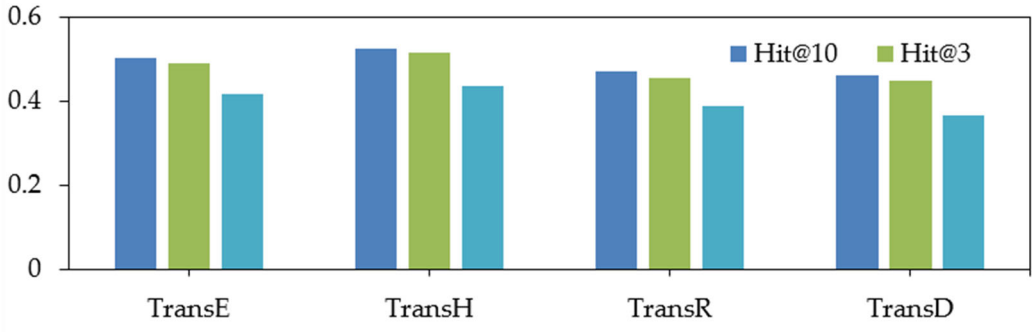


Figure 10. Comparison results of knowledge representation models.

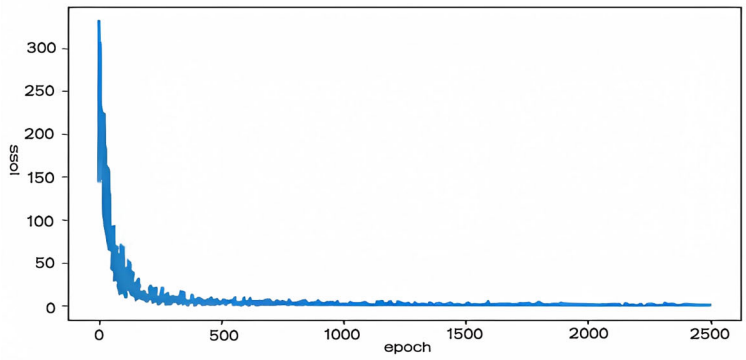


Figure 11. The changing trend of TransH model loss function.

5. Conclusions

In order to extract knowledge from unrelated, heterogeneous, multi-source data, knowledge extraction methods were investigated and a professional knowledge base of potato diseases and pests was constructed. Knowledge representation methods were tested to represent the constructed knowledge graph for further application. The main summary of the work is as follows:

(1) Dataset construction with crawling and digital technology was used to obtain a large amount of semi-structured and unstructured data on potato pests and diseases from agricultural specialist websites and books. The collected data was pre-processed with cleaning and the removal of redundancies to form a potato pest and disease data corpus, PotatoRE;

(2) Based on the PotatoRE corpus, the entity types and relationships of potato pests and diseases were defined under the guidance of crop protection experts. Three annotation methods, BIO, BMES and BIOES, were compared on the PotatoRE corpus. The better annotation method, BMES, was selected to annotate the corpus and form a high-quality dataset consisting of six entity types, eight relationships and four attribute types, totalling 8971 entity samples, for subsequent experiments;

(3) An ALBert-BiLSTM-Self_Att-CRF model was developed to extract entities and relationships from the data corpus and construct a domain knowledge graph of potato diseases and pests. To verify the efficiency and accuracy of the proposed model, it was experimentally compared with eight other SOTA models on the dataset constructed in this paper. The results show that the model designed in this paper performed well in terms of the accuracy, recall, and F1 values. Compared to the Al-Bert-BiLSTM-CRF and ALBert-BiGRU-CRF models, their accuracy was improved by 2.92% and 3.12%, respectively. Compared with the Bert-BiLSTM-CRF and Bert-BiGRU-CRF models, the model in this paper not only improved the accuracy, recall, and F1 values, but also saved training time and improved the efficiency of entity and relationship extraction. In addition, the robustness of the proposed model was further verified by comparing it with other mainstream models using the People's Daily dataset;

(4) The performance of the knowledge representation models TransE, TransH, TransR, and TransD was compared experimentally using the constructed knowledge graph. The better model, TransH, was then used to represent the knowledge graph in this paper, which laid the foundation for knowledge inference and fusion and enhanced its application value.

The shortcoming of the model in this paper is that annotating the data takes a lot of time. Future research will focus on automatic annotation methods or using LLM to extract knowledge from multimodal data to improve the efficiency of knowledge base construction and practical applications.

Author Contributions: Conceptualization, W.Y. (Wanxia Yang); methodology, W.Y. (Wanxia Yang) and S.Y.; software, S.Y. and Y.L.; formal analysis, W.Y. (Wanxia Yang) and S.Y.; investigation, W.Y. (Wanxia Yang) and G.W.; data curation, S.Y.; writing—original draft preparation, W.Y. (Wanxia Yang), and G.W.; writing—review and editing, Y.L., J.L. and W.Y. (Weiwei Yuan). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Gansu Province University Youth Doctoral Fund Project (Project number and name: 2021QB-033, Research on the Construction Method and Application of Knowledge Graph for the Whole Potato Industry Chain in Gansu Province).

Data Availability Statement: Presented data in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflict of interest.

References

1. Guo, X.; Zhou, H.; Su, J.; Xia, H.; Li, L. Chinese agricultural diseases and pests named entity recognition with multi-scale local context features and self-attention mechanism. *Comput. Electron. Agric.* **2020**, *179*, 105830. [CrossRef]
2. Xia, Y.; Sun, N.; Wang, H.; Yuan, X.; Wang, C.; Gao, Q. Research on knowledge question answering system for agriculture disease and pests based on knowledge graph. *J. Nonlinear Convex Anal.* **2020**, *21*, 1487–1496.
3. Kurmi, Y.; Gangwar, S.; Agrawal, D.; Kumar, S.; Shanker, H. Leaf image analysis-based crop diseases classification. *Signal Image Video Process.* **2021**, *15*, 589–597. [CrossRef]
4. Zhao, C. Agricultural Knowledge Intelligent Service Technology: A Review. *Smart Agric.* **2023**, *5*, 126–148.
5. Zhang, M.; Yang, Z.; Liu, C.; Fang, L. Traditional Chinese Medicine knowledge Service based on Semi-Supervised BERT-BiLSTM-CRF Model. In Proceedings of the 2020 International Conference on Service Science (ICSS), Xining, China, 24–26 August 2020.
6. Sun, Y.; Wang, S.; Li, Y.; Feng, S.; Chen, X.; Zhang, H.; Tian, X. ERNIE: Enhanced Representation through Knowledge Integration. *arXiv* **2019**, arXiv:1904.09223.
7. Chen, J.; Xi, X.; Pi, Z.; Sheng, S.; Cui, Z. ALBERT-Based Named Entity Recognition of Chinese Medical Records. *J. Nanjing Norm. Univ. (Engl. Technol. Ed.)* **2021**, *21*, 36–43. [CrossRef]
8. Wu, C.; Luo, G.; Guo, C.; Yi, R.; Zhen, A.; Yang, C. An Attention-based Multi-Task Model for Named Entity Recognition and Intent Analysis of Chinese Online Medical Questions. *J. Biomed. Inform.* **2020**, *108*, 103511. [CrossRef] [PubMed]
9. Zhang, B.; Liu, K.; Wang, H.; Li, M.; Pan, J. Chinese named-entity recognition via self-attention mechanism and position-aware influence propagation embedding. *Data Knowl. Eng.* **2022**, *139*, 101983. [CrossRef]
10. Taher, E.; Hoseini, S.A.; Shamsfard, M. Beheshti-ner: Persian named entity recognition using bert. *arXiv* **2020**, arXiv:2003.08875.
11. Zhang, N.; Xu, G.; Zhang, Z.; Li, F. Mifm: Multi-granularity information fusion model for chinese named entity recognition. *IEEE Access* **2019**, *7*, 181648–181655. [CrossRef]
12. Zhang, Q.; Sun, Y.; Zhang, L.; Jiao, Y.; Yue, T. Named entity recognition method in health preserving field based on bert. *Procedia Comput. Sci.* **2021**, *183*, 212–220. [CrossRef]
13. Hakala, K.; Pyysalo, S. Biomedical Named Entity Recognition with Multilingual BERT. In Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, Hong Kong, China, 4–6 November 2019; pp. 56–61.
14. Zhao, P.; Zhao, C.; Wu, H.; Wang, W. Multi-feature fusion agricultural named entity recognition based on BERT. *Trans. Chin. Soc. Agric. Mach.* **2022**, *38*, 112–118.
15. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBert: A lite bert for self-supervised learning of language representations. *arXiv* **2019**, arXiv:1909.11942.
16. Zhao, P.; Zhao, C.; Wu, H. Named entity recognition of Chinese agricultural text based on attention mechanism. *Trans. Chin. Soc. Agric. Mach.* **2021**, *52*, 185–192.
17. Guo, X.; Hao, X.; Tang, Z.; Diao, L.; Bai, Z.; Lu, S. ACE-ADP: Adversarial contextual embeddings based named entity recognition for agricultural diseases and pests. *Agriculture* **2021**, *11*, 912. [CrossRef]

18. Socher, R.; Chen, D.; Manning, C.D.; Ng, A.Y. Reasoning with Neural Tensor Networks for Knowledge Base Completion. In Proceedings of the NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems, Lake Tahoe Nevada, CA, USA, 5–10 December 2013; Volume 1, pp. 926–934.
19. Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; Yakhnenko, O. Translating Embeddings for Modeling Multi-relational Data. In Proceedings of the NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems, Lake Tahoe Nevada, CA, USA, 5–10 December 2013; Volume 2, pp. 2787–2795.
20. Wang, Z.; Zhang, J.; Feng, J.; Zheng, C. Knowledge Graph Embedding by Translating on Hyperplanes. In Proceedings of the AAAI'14: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, Québec City, QC, Canada, 27–31 July 2014; pp. 1112–1119.
21. Moon, C.; Jones, P.; Samatova, N. Learning entity type embeddings for knowledge graph completion. In Proceedings of the CIKM'17: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, New York, NY, USA, 6–10 November 2017; pp. 2215–2218.
22. Ji, G.; He, S.; Xu, L.; Liu, K.; Zhao, J. Knowledge Graph Embedding via Dynamic Mapping Matrix. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 26–31 July 2015; pp. 687–696.
23. Tang, B.; Wang, X.; Yan, J.; Chen, Q. Entity recognition in Chinese clinical text using attention-based CNN-LSTM-CRF. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 89–97. [CrossRef] [PubMed]
24. Lin, Y.; Liu, Z.; Luan, H.; Sun, M.; Rao, S.; Liu, S. Modeling relation paths for representation learning of knowledge bases. *arXiv* **2015**, arXiv:1506.00379.
25. Yang, J.; Zhang, Y.; Li, L.; Li, X. Yedda: A lightweight collaborative text span annotation tool. *arXiv* **2018**, arXiv:1711.03759.
26. Arora, S.; Li, Y.; Liang, Y.; Ma, T.; Risteski, A. Linear algebraic structure of word senses, with applications to polysemy. *arXiv* **2018**, arXiv:1601.03764. [CrossRef]
27. Zhang, W.; Jiang, S.; Zhao, S.; Hou, K.; Zhang, L. A BERT-BiLSTM-CRF Model for Chinese Electronic Medical Records Named Entity Recognition. In Proceedings of the 2019 12th International Conference on Intelligent Computation Technology and Automation (ICICTA), Xiangtan, China, 26–27 October 2019.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.