information

MDPI

Article

Automated Knowledge Extraction in the Field of Wheat Sharp Eyespot Control

Keyi Liu [1,2] and Yunpeng Cui [1,2,*]

[1] Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, China; liukeyi@caas.cn
[2] Key Laboratory of Big Agri-Data, Ministry of Agriculture and Rural Areas, Beijing 100081, China
* Correspondence: cuiyunpeng@caas.cn

Abstract: Wheat sharp eyespot is a soil-borne fungal disease commonly found in wheat areas in China, which can occur throughout the entire reproductive period of wheat and has a great impact on the yield and quality of wheat in China. By constructing a domain ontology for wheat sharp eyespot control and modeling the domain knowledge, we aim to integrate and share the knowledge in the field of wheat sharp eyespot control, which can provide important support and guidance for agricultural decision-making and disease control. In this study, the literature in the field of wheat sharp eyespot control was used as a data source, the KeyBERT keyword extraction algorithm was used to mine the core concepts of the ontology, and the hierarchical relationships among the ontology concepts were extracted through clustering. Based on the constructed ontology of wheat sharp eyespot control, the schema of knowledge extraction was formed, and the knowledge extraction model was trained using the ERNIE 3.0 knowledge enhancement pretraining model. This study proposes a model and algorithm to realize knowledge extraction based on domain ontology, describes the construction method and process framework of wheat sharp eyespot control domain ontology, and details the training and reasoning effect of the knowledge extraction model. The knowledge extraction model constructed in this study for wheat sharp eyespot control contains a more complete conceptual system of wheat sharp eyespot. The F1 value of the model reaches 91.26%, which is a 17.86% improvement compared with the baseline model, and it can satisfy the knowledge extraction needs in the field of wheat sharp eyespot control. This study can provide a reference for domain knowledge extraction and provide strong support for knowledge discovery and downstream applications such as intelligent Q&A and intelligent recommendation in the field of wheat sharp eyespot control.

Keywords: wheat sharp eyespot; knowledge extraction; domain ontology; automatic extraction

Citation: Liu, K.; Cui, Y. Automated Knowledge Extraction in the Field of Wheat Sharp Eyespot Control. *Information* **2024**, *15*, 367. https://doi.org/10.3390/info15070367

Academic Editor: Domenico Fabio Savo

Received: 24 April 2024
Revised: 23 May 2024
Accepted: 20 June 2024
Published: 21 June 2024

1. Introduction

Wheat sharp eyespot (*Rhizotonia cerealis* van der Hoeven apud.Boerema & Verhoeven) is a worldwide soil-borne fungal disease caused by *Rhizoctonia cerealis* van der Hoeven [1,2]. At present, the disease occurs to varying degrees in nearly twenty provinces and cities in China, especially in the wheat areas of Jiangsu, Zhejiang, Anhui, Shandong, Henan, Hebei, Shaanxi, Guizhou, Hubei, and Sichuan provinces, where it is more prevalent. Wheat sharp eyespot can occur throughout the entire reproductive period of wheat, causing a variety of symptoms such as rotting buds, diseased and dead seedlings, rotting stems of flower stalks, and collapse [3]. In recent years, the incidence of blight in straw-returned wheat areas in northern China has been increasing year by year, seriously jeopardizing the yield and quality of wheat. Wheat sharp eyespot has a huge impact on yield, usually leading to a reduction in wheat yield between 10% and 20%, and in severe cases, even up to about 50% reduction in yield, and even complete crop failure in individual plots [4]. Farmers in the process of wheat disease control mostly rely on experience. Although many farmers are already experts in wheat diseases, the level of knowledge varies, and

regional differences, climate differences, and other factors pose challenges for accurate disease control. Mostly relying on past experience to apply pesticides, the issue of grain quality and safety is ignored in the pursuit of yield. Although scientific and technological research is an important carrier of technological innovation, it is difficult for farmers to gain knowledge of wheat sharp eyespot in the literature, and it is difficult to judge the authenticity of the relevant knowledge through the Internet. Therefore, by constructing an ontology for wheat sharp eyespot, modeling the domain knowledge, and extracting the knowledge in the scientific and technological literature based on the knowledge model of wheat sharp eyespot, we can provide more scientific guidance to the producers so that farmers can understand how to control wheat sharp eyespot according to the onset of the disease and can independently use the professional methods of wheat sharp eyespot prevention and control, improve the efficiency of wheat disease prevention and control, and reduce the losses caused by the disease. It is of great significance to China's agricultural production and food security.

Massive literature data bring great challenges to technical analysis. As an unstructured text, the literature is not uniform in description or terminology, which makes it difficult to use simple rules to extract the core knowledge in it, and the current extraction method of manual participation in annotation can no longer meet the needs of rapid analysis of large-scale datasets. Entity–relationship extraction is one of the important tasks of information extraction, aiming at extracting the semantic relationships between different entities from unstructured text so as to extract useful information. In 1998, the first Message Understanding Conference (MCU) was held on entity–relationship extraction, and after years of development, the general field of relationship extraction has significantly advanced. Users use the information extraction method to obtain the required information from a large number of data sources, and on this basis, the extracted information is processed, organized, and analyzed through a secondary approach to obtain a new understanding of the information, which can help to improve the understanding of the original information and then reach the level of knowledge.

This study is oriented toward the field of wheat sharp eyespot control, aiming at integrating the relevant knowledge in this field through knowledge modeling, constructing an ontology of wheat sharp eyespot control that is useful for agricultural decision-making and practice in response to the actual needs in this field, and applying the domain ontology to the knowledge extraction process in order to improve the accuracy and efficiency of the extraction, so as to provide important support and guidance for agricultural decision-making and disease control. At the same time, the methods and results of this study will also provide reference and inspiration for knowledge extraction in other fields.

Our major contributions are the following:

- We propose a method for constructing an ontology in the field of wheat sharp eyespot control, detailing the process framework for building the domain ontology based on a corpus of wheat sharp eyespot research, which facilitates the integration and sharing of knowledge in this field.
- Based on the ontology of wheat sharp eyespot control, we introduce a knowledge extraction model specific to this domain, forming a framework for knowledge extraction that effectively extracts relevant information about wheat sharp eyespot control from texts.
- The knowledge extraction model and algorithm we proposed have achieved significant results in the field of wheat sharp eyespot control and also provide a reference for knowledge extraction in other domains.

The structure of the paper is as follows: In Section 2, we discuss the background concepts and overview of the research methodology. Section 3 describes the research methodology for datasets, ontology construction, and automated knowledge extraction. Section 4 presents the current limitations and future research directions, and the conclusion is presented in Section 5.

2. Related Work

2.1. Ontology

In the 1980s, ontology was introduced into the field of information science and later gradually extended to the fields of knowledge engineering and artificial intelligence [5]. In 2001, the Food and Agriculture Organization of the United Nations (FAO) introduced the Agricultural Ontology Service (AOS), which provides users with a way to define and describe domain knowledge through an ontological methodology and assists communication within the domain. Based on the AOS project, crop ontology, food ontology, and agronomy ontology were constructed [6,7]. Chang Chun (2003) introduced the AOS into the Chinese context, which was the beginning of agricultural ontology research in China [8]. Since then, many scholars have started to elaborate on the construction principles, processes, methods, and other contents of ontologies in the field of agriculture. In addition to the theoretical study of ontology, it has also been applied and practiced in many aspects of the agricultural field, such as crop production and cultivation, pest control, germplasm resources, and so on. Among them, the ontology research of crop pest control is the most extensive, considering areas such as rice pests and diseases [9,10], maize pests and diseases [11,12], and cotton diseases [13]. In the field of agriculture, a commonly used ontology construction method is semi-automatic construction, by transforming agricultural narrative lists and knowledge categorization into ontologies. Currently, the commonly used agricultural narrative lists are the Chinese Agricultural Thesaurus (CAT), AGROVOC (Multilingual Agricultural Thesaurus), NALT (National Agricultural Library of the United States), and the Centre Agriculture Bioscience International (CABI). Liu Guifeng et al. (2022) used the National Agricultural Science Data Center's Cotton, Hemp Crop Pathogenic Fungal Disease Database, and Microbial Pesticide Database as their main data sources and combined the classification and terminology in the Chinese Classification Thesaurus, Agricultural Science Narrative Thesaurus, Chinese Thesaurus, and Chinese Library Classification to construct the "Cotton Disease Prevention and Control" ontology [13]. Renny et al. (2021) used text mining and machine learning to construct an ontology for the tomato pest and disease domain [14]. Deepa R et al. (2022) used natural speech processing techniques to extract agricultural terms and combined textual similarity with plain Bayes (NBM) to propose a method for automatically constructing agricultural ontologies [15]. In specific verticals, narrative lists are difficult to provide finer-grained knowledge categorization, and ontology construction relies more on domain experts by manual construction. Dong et al. (2023) used a top-down modeling approach, in which domain experts manually compiled concepts from the topmost level, and agricultural experts further refined the relationships and hierarchies among knowledge to construct an ontology model for precision rice fertilization [16]. Agricultural ontology research has made some progress in China, especially in crop pest control, which is widely used. However, there are some shortcomings, including the inadequacy of narrative lists for fine-grained classification, subjective factors in the semi-automatic construction process, challenges in updating and maintaining ontologies, lack of standardization, and complexity of expertise acquisition and validation. In order to improve the quality and sustainability of agricultural ontology research, there is a need to explore more effective construction methods and promote standardization.

2.2. Knowledge Extraction

In response to all the above problems arising from the pipeline approach, scholars gradually began to study the strengthening of the link between two subtasks. As a result, joint extraction methods have been proposed and have been continuously developed in recent years. The joint entity–relationship extraction method can further utilize entity recognition and relationship to extract the potential information between two subtasks, which often achieves better results than the pipeline method. Entity–relationship extraction is the joint completion of entity recognition and relationship extraction, directly from the text to obtain the entity–relationship ternary. Its mathematical description is as follows: The set of relationship categories R and the set of entity categories E are known. Given

the sentence S = {w_1, w_2, ⋯, w_n}, using the entity–relationship joint extraction method, all entity–relationship quintuples <*h*, e_1, *r*, *t*, e_2> in the sentence S are extracted by using the established unified model, where $r \in R$, $e_1 \in E$, and $e_2 \in E$. The significant difference between this method and the streamlined approach is that there is no prelabeling of the given entity boundaries and types, and the joint model outputs all the relation triples <*h*, *r*, *t*> for sentence S.

Zeng et al. [17] designed CNNs for relationship classification but did not consider a ternary form. Makoto et al. [18] achieved end-to-end prediction of entity relationships by constructing a stacked network based on BiLSTM and Bi-TreeLSTM for both entity extraction and relationship detection. Li et al. [19] used a two-layer LSTM with an encoder–decoder architecture to construct a knowledge extraction model that is not limited to the ternary form and can predict structured knowledge in a fixed format. Zheng et al. [20] transformed the entity–relationship extraction task into a sequence annotation task through an annotation strategy and then constructed a Bi-LSTM model similar to the previous one to handle it. This enables the model to extract triples directly from statements, but this model also has a design flaw, namely that it cannot deal with data with overlapped triples. Luan et al. [21] designed a multi-task learning framework for recognizing entities and relationships in scientific research to construct a scientific knowledge graph, and this model outperforms the existing models without any prior knowledge of the domain. Wang et al. [22] presented a new handshake labeling strategy to label the entity head to entity tail, subject head to object head, and subject tail to object tail, to decompose the joint extraction task into sequence labeling subtasks in order to solve the problem of exposing bias in the training and prediction phases. Sui et al. [23] used a transformer as a decoder, stacked multiple identical transformers, and used a multi-head self-attention mechanism to simulate the relationship between the triples, fusing sentence information and giving an attentional representation to the sentence through multi-head mutual attention.

Although the above works have achieved good results and pushed forward the progress and development of the field of joint entity–relationship extraction, none of them focuses on how to fully mine semantic and syntactic information, or how to reasonably fuse the two kinds of information for different contexts. Classical relational extraction tends to focus on a single sentence and just tries to mine the entity relationships within each sentence. Early research in relational extraction focused on categorizing relationships between pairs of entities in a single sentence or jointly extracting entities and relationships in a sentence, i.e., intra-sentence relationships, ignoring the relationships of pairs of entities that cross sentence boundaries, i.e., inter-sentence relationships. Therefore, entity–relationship extraction is inevitably limited to a certain extent in practical application scenarios. In reality, a large number of relationships actually need to rely on cross-sentence or even document linguistic information in order to be extracted; this document will mention many modern composite cross-passing-related system entities, where a composite of multiple sentence scenarios is proposed according to which the relationship needs are read, memorized, and reasoned in order to find out the relationship facts between multiple sentences. There are many relationship facts that are hidden in entity pairs of different sentences in a document, and there are complex interactions between multiple entities in a document. With the continuous progress of technology and the deepening of the research on relationship extraction, the demand for document-level entity–relationship extraction is rapidly increasing.

Ontology-based information extraction techniques, with the help of predefined ontology hierarchies, can effectively identify domain-specific concepts, entities, relationships, and other forms of knowledge. Ontology can be regarded as a tree-structured knowledge base mold, which is the semantic basis for communication and connectivity between different subjects in the same domain. Moreno [24] proposed a method to achieve information extraction based on ontology in an independent domain. The application was oriented toward the field of molecular biology, the extraction of information about E. coli, and the establishment of a regulatory network for E. coli. The constructed system was tested on

the abstracts of scientific papers and the complete literature in this field. Additionally, the complete literature was mined by designing the domain ontology and then implementing information extraction based on the knowledge contained in the ontology. Li et al. [25] implemented an agricultural domain based on agricultural ontology for the extraction of structured AJAX data. Daya [26] used multiple ontologies for information extraction considering the two cases of subdomain determination and subdomain expression. The first system based on multiple ontologies was developed for the university domain, which uses two ontologies specialized in subdomains. The corpus consists of documents from 100 universities, 50 web pages from North America, and 50 from the rest of the world. The second system realized was applied to the domain of terrorist attacks, and the corpus used by the Message Understanding Conference (MUC) provides the subdomains.

3. Method

3.1. Dataset Construction

In this paper, the literature in the field of wheat sharp eyespot control was collected using the CNKI as a data source, which allowed us to obtain more concepts and relationships compared to structured data, making the constructed ontology of wheat sharp eyespot control more comprehensive and complete. Scientific and technical research often comes with fixed keywords, but the number of keywords is not enough to characterize the core concepts of the field, and keyword extraction for chapter-level documents can reveal more comprehensive concepts. Using "Wheat Sharp Eyespot AND (control OR prevention)" as the search formula, a total of 1008 documents were found. The corpus of wheat sharp eyespot control was obtained by formatting the documents, preprocessing the text, and filtering the words by word splitting and stopword filtering.

3.2. Ontology Construction

The skeletal methodology, the TOVE method, the methontology method, the five-step cycle method, and the seven-step method are the typical ontology modeling methods available. The skeletal methodology was summarized by Mike Uschold and King from enterprise ontology construction at the University of Edinburgh in 1995. The TOVE methodology was proposed by Gruninger and Mark S. Fox from the Enterprise Integration Laboratory of the University of Toronto in 1996 during the TOVE project. The methontology method was proposed by Mariano Fernandez and Gomez Gomez from the University of Madrid Crafts Campus. Mariano Fernandez and Gomez Perez et al. developed in 1997 [27]. The five-step cycle method was developed for the construction of semantic web ontology learning by Maedche and Staab in 2000. The seven-step method is a domain ontology construction method developed by Noy and McGuinness at Stanford University School of Medicine in 2000. A comparison of the characteristics and uses of these five methods is shown in Table 1.

Table 1. Ontology construction methods.

Ontology Modeling Methods	Application Areas	Basic Processes	Drawbacks	Life Cycle	Reusable or Not
Skeletal Methodology	Corporate Area	Defining the purpose and scope of ontology applications; Building ontologies; Evaluation; Documentation	Lack of specific methodologies and techniques	No life cycle	no
TOVE	Corporate Area	Clarify the purpose of the construction; Formulate the methodology; Formalize the steps; Constraints; Test and revise the ontology	Lack of documented process descriptions and specific build steps	No true life cycle	no
Methontology	Chemical Field	Specification; Knowledge acquisition; Conceptualization; Integration; Realization; Evaluation; Documentation	Unable to update iterations	Life cycle	no

Table 1. *Cont.*

Ontology Modeling Methods	Application Areas	Basic Processes	Drawbacks	Life Cycle	Reusable or Not
Five-Step Cycle	Semantic Web Ontology Learning	Ontology import and reuse; Ontology extraction; Ontology trimming; Ontology refinement; Ontology application	Poorly operated and difficult	Life cycle	yes
Seven-Step Process	Medical Field	Define domain scope; Reuse existing ontologies; List conceptual terms; Define classes and inter-class hierarchies; Define class attributes; Define facets of attributes; Create instances	Lack of ontology assessment to update iterations	No true life cycle	yes

The ontology construction process is proposed to meet the actual needs of wheat sharp eyespot control (Figure 1). The topic of "wheat sharp eyespot control" is systematic, covering a wide range of fields such as agricultural science, plant protection, chemistry, biology, pesticide science, etc. The knowledge related to wheat is extremely large. This makes it impossible to build an ontology that covers all of it. Therefore, the focus of this study is to construct the ontology of "wheat sharp eyespot control", eliminate irrelevant subconcepts, retain the overall framework structure, and on this basis, add new concepts and their subconcepts to realize the integration, supplementation, and improvement of the ontology concepts of wheat sharp eyespot control and develop the conceptual framework structure.

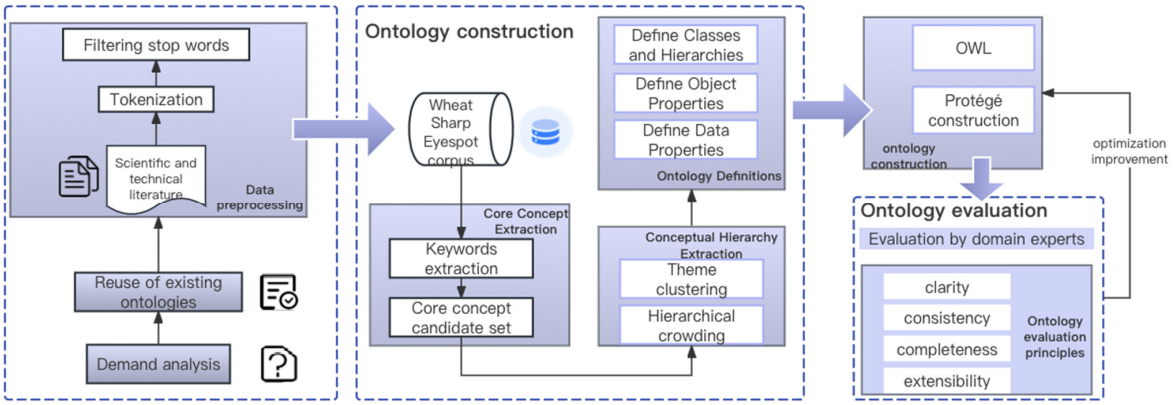

Figure 1. Ontology construction process.

In this study, the KeyBERT keyword extraction algorithm was used to find the most similar subphrases in a document to the document itself using BERT embedding and cosine similarity. Firstly, BERT was used to calculate the embedding value of the document to obtain a vector-level representation of the document. Then, the word vectors were extracted for the n-gram, and finally, cosine similarity was used to determine the keywords or key phrases that were most similar to the document to obtain the keywords that best described the whole document. In order to diversify the results, we used maximum marginal relevance (*MMR*) to create keywords/key phrases, also based on cosine similarity. The specific formula is as follows:

$$MMR(Q, C, R) = \underset{d_i \in C}{\operatorname{argmax}}[\lambda sim(Q, d_i) - (1 - \lambda) \max_{d_j \in R}^{k} (sim(d_i, d_j))] \quad (1)$$

where Q denotes the query statement; C denotes the set of all documents; R denotes an initial set that has been obtained based on the relevance; $\arg\max^{k}[*]$ denotes the index

that gives the k largest elements of the set; $sim(Q, d_i)$ represents the correlation between d_i and Q; $sim(d_i, d_j)$ represents the redundancy of the representation.

On the basis of the keywords, they were further screened and condensed as a core concept candidate set, taking into account the ontological needs of wheat sharp eyespot control. Table 2 demonstrates five sets of keywords.

Table 2. Keywords.

Keywords
[('Disease-resistant, Varieties, Wheat, Varieties', 0.7423), ('Wheat, Variety, Field, Resistance', 0.7211), ('Fertility, Disease prevention, Wheat, Population', 0.7206), ('Yumai, New wheat, Variety, Resistance', 0.7189), (Disease prevention, Wheat, Population, Structure', 0.717)]
[('Disease Strain, Disease Strain, Carrying Bacteria, Overwintering', 0.6457), ('Wheat fields, Morbidity, Climate', 0.594), ('Spring, Early, Onset, Sources of Infestation', 0.5845), ('Wheat, Wheat Sharp Eyespot, Period of occurrence', 0.5814), ('Carrying bacteria, Overwintering, Next year, Spring', 0.5766)]
[('Temperature, Humidity, Wheat, Wheat Sharp Eyespot', 0.6713), ('Spring, High humidity, First third of a month, Rainfall', 0.6434), ('Humidity, Wheat, Wheat Sharp Eyespot, Occurrence', 0.6238), ('Humidity, Wheat, Wheat Sharp Eyespot', 0.6233), ('Wheat Sharp Eyespot, Research, Hot Spots ', 0.538)]
[('Climate, Conditions, Wheat, Wheat Sharp Eyespot', 0.6904), ('Agriculture, Prevention and Control, Suzhou City, Climate', 0.6856), ('Climate, Soil, Growth, Wheat', 0.6569), ('Prevention and Control, Suzhou City, Climate, Soil', 0.6548), ('Disease, Resistance, Control, Wheat Field', 0.6446)]
[('Field, Morbidity, Overwintering, Fertilization', 0.6605), ('Infestation, Bacterial source, Wheat, Sowing', 0.643), ('Incidence, Seeding rate, Wheat field, Seasonal period', 0.6325), ('Field, Pathogen, Quantity, Deep plowing', 0.6318), ('Overwintering, Initial infestation, Bacterial source', 0.6144)]

After clarifying the domain scope and the core concept candidate set, the first task of constructing a domain ontology is to model the ontology hierarchy. In this study, we used BERTopic for topic modeling, which uses Transformer and c-TF-IDF to create dense clusters that can obtain topics and important words. The main steps are as follows: First, input the whole document into BERT to obtain the word vectors of the document; apply UMAP to downsize these word vectors to obtain the low-dimensional word vectors; cluster the low-dimensional word vectors to obtain the documents with good clustering classes; use c-TF-IDF to obtain the subject words of each topic for the clustered documents; use maximum marginalization for these subject words; and use c-TF-IDF to obtain the subject words of each topic. These words were filtered using the maximum marginal correlation algorithm. By clustering the core concept candidate set of topics, 50 topics were obtained. Table 3 shows the top five topics. Topic "−1" is the largest, which refers to the outlier text that is not assigned to any of the generated topics and will be ignored in this paper.

Table 3. Topic clustering.

Topic	Count	Name	Representation
−1	1787	−1_Wheat_control_Wheat Sharp Eyespot _occurrence	['wheat', 'control', 'Wheat Sharp Eyespot', 'occurrence', 'disease', 'incidence', 'varieties', 'agents', 'research', 'impact', 'agriculture', 'trials', 'growth', 'seed', 'efficacy', 'seed mixes', 'technology', 'field', 'survey', 'soil']
0	177	0_Gene_Marker_Resistance_Inheritance	['gene', 'marker', 'resistance', 'genetic', 'chromosome', 'identification', 'analysis', 'localization', 'detection', 'population', 'expression', 'molecular', 'research', 'Wheat Sharp Eyespot', 'chain', 'material', 'trait', 'mapping', 'protein', 'utilization']
1	127	1_Tests_Pharmaceuticals_Investigations_Effectiveness	['test', 'agent', 'investigation', 'efficacy', 'plot', 'Wheat Sharp Eyespot', 'wheat', 'application', 'seed dressing', 'seed', 'seed coating', 'control', 'control efficacy', 'ltd', 'penicillin', 'method', 'phenoxyethanol', 'suspension', 'for test', 'year/month/day']
2	111	2_Occurrence_Wheat_Wheat Sharp Eyespot_Disease	['occurrence', 'wheat', 'Wheat Sharp Eyespot', 'onset', 'disease', 'area', 'plant', 'average', 'variety', 'infestation', 'control', 'survey', 'damage', 'million acres', 'impact', 'disease', 'leaf sheath', 'corn', 'extent', 'symptoms']

Table 3. *Cont.*

Topic	Count	Name	Representation
3	101	3_Research_Journal_of_Wheat Sharp Eyespot_Strains	['Research', 'Journal', 'Wheat Sharp Eyespot', 'Strain', 'Bacteria', 'Wheat', 'Agriculture', 'Plant', 'China', 'Nucleobacteria', 'Screening', 'Science', 'Isolation', 'Identification', 'Beijing', 'Bioprophylaxis', 'Antagonism', 'Prevention and control', 'Publisher', 'Henan']
4	97	4_Sowing_Control_Wheat_Soil	['sowing', 'control', 'wheat', 'soil', 'field', 'wheat field', 'Wheat Sharp Eyespot', 'nitrogen fertilizer', 'fertilization', 'mulching', 'raising', 'deep loosening', 'lowering', 'potash', 'impact', 'incidence', 'control', 'weeds', 'occurrences', 'kilogram']
5	93	5_Disease_leaf sheaths_wheat_leaf blades	['onset', 'leaf sheath', 'wheat', 'leaf', 'disease', 'infestation', 'basal', 'control', 'spot', 'occurrence', 'brown', 'symptom', 'dieback', 'plant', 'stalk', 'disease', 'sowing', 'diseased plant', 'seed', 'white spike']

3.2.1. Wheat Sharp Eyespot Control Ontology Concept Definition

In this study, the results of theme and hierarchical clustering were analyzed and condensed, and the ontology of wheat sharp eyespot control was divided into eight categories of parent concepts, namely incidence characteristics, wheat growing period, pathogenesis, region of incidence, degree of disease, lesion site, symptom, and control measures. In accordance with the structural requirements of OWL, they are all subordinate concepts of "Thing". The Protégé selected in this article is a widely used open source ontology editor developed by Stanford University. It is based on Java and is used for ontology editing and knowledge acquisition. Protégé has an extensible structure and multiple plug-ins; supports ontology description languages such as RDF, RDFS, and OWL; and provides a graphical interface that is easy to use.

Using Protégé, we incorporated hierarchical concepts into the "Classes" based on data content and representation characteristics. This was accomplished in collaboration with domain experts and in accordance with ontology design principles, ensuring a balanced scope of knowledge description and conceptual refinement (Figure 2).

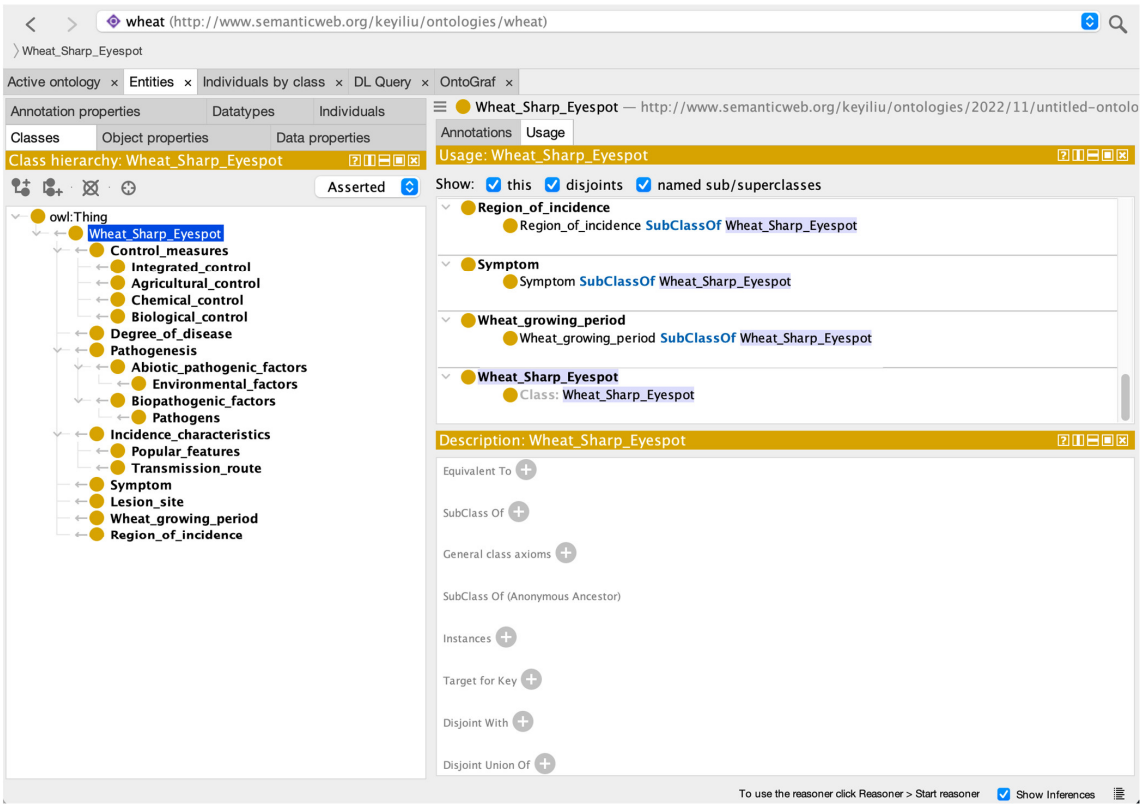

Figure 2. Ontology concept.

3.2.2. Wheat Sharp Eyespot Control Ontology Attribute Definition

Once the conceptual hierarchy of an ontology is constructed, the relationships between concepts need to be determined, i.e., defining ontology property relationships [28]. Ontology property relationships are the basis for the subsequent implementation of knowledge reasoning. Three types of property relationships are covered in the Protégé tool, namely object properties, data properties, and annotation properties. Object properties are used to represent the relationships between concepts, data properties are used to describe the properties of the concepts themselves, and annotation properties are used to annotate the concept properties. In this study, we considered only object properties and data properties. According to the characteristics of the field of wheat sharp eyespot control, by organizing and analyzing the associations between parent concepts, the object attribute relationships involved in the ontology are shown in Table 4. In this study, 11 object attributes, 16 primary data attributes, and 8 secondary data attributes were defined for the wheat sharp eyespot control ontology. The object properties and data properties options in Protégé were used to add object properties and data properties, respectively (Figure 3). The characteristic information of each object property was used to define its nature. Then, the corresponding description information was filled, and constraint information on aspects such as definition domain, value domain, and inverse property was added [28].

Table 4. Object attribute definition.

Type	Attribute Relationship	Relationship Description	Domain	Range	Reciprocal Attribute
Object Properties	beCausedBy	Caused by . . .	Wheat Sharp Eyespot	Environmental factors, pathogens	cause
	beControledBy	Controlled by . . .	Wheat Sharp Eyespot	Control measures	control
	cause	Lead to	Environmental factors, pathogens	Wheat Sharp Eyespot	beCausedBy
	control	Relationship between prevention and control	Control measures	Wheat Sharp Eyespot	beControledBy
	harmOn	Harm relationship	Wheat Sharp Eyespot	Lesion site	none
	hasChannel	Transmission route	Wheat Sharp Eyespot	Transmission route	none
	hasCharacteristics	Popular features	Wheat Sharp Eyespot	Popular features	none
	hasSymptom	Symptomatic	Diseased or infected plant	Symptom	none
	infectPeriod	Disease Infestation Stage	Wheat Sharp Eyespot	Wheat growing period	none
	occurredIn	Disease Areas	Wheat Sharp Eyespot	Region of incidence	none

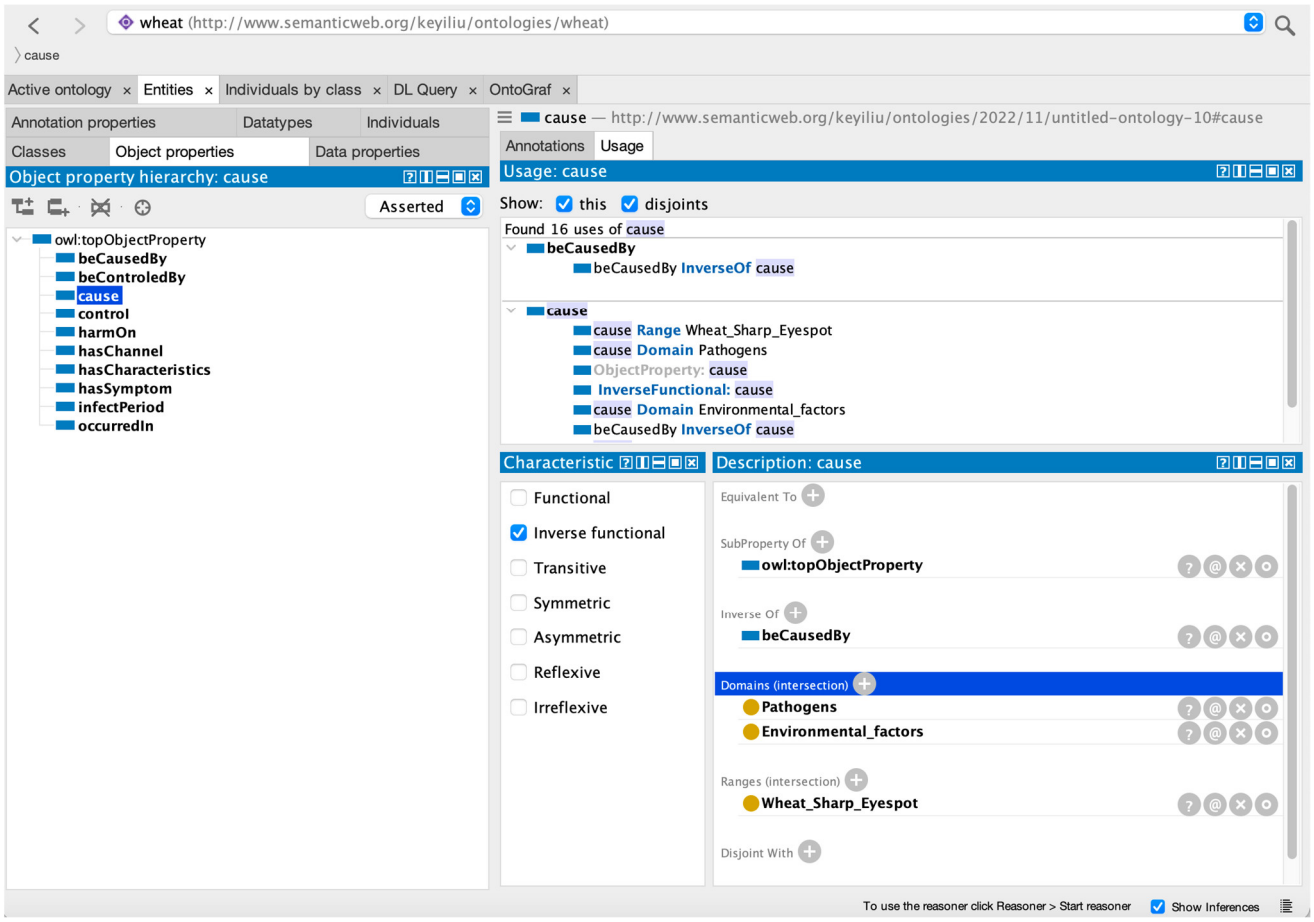

Figure 3. Object properties and data properties.

3.2.3. Ontology Construction and Evaluation

In this study, the constructed ontology was evaluated and optimized according to the ontology evaluation criteria to ensure the scientific and professional nature of the ontology so that it can fully express the concepts of the wheat sharp eyespot control domain. The criteria for ontology evaluation usually include clarity, consistency, refinement, and extensibility [29,30]. Clarity requires that the defined classes and attributes must be clear and free of ambiguity; consistency requires that the relationships between classes are logically consistent [30]; refinement requires that the defined classes and attributes can describe the wheat sharp eyespot control domain in a complete way and be applicable to the body of knowledge in the main data sources; and extensibility requires that the ontology can be extended in the wheat sharp eyespot control domain to accommodate the emergence of new concepts. After the initial construction of the ontology was completed, experts in the field of plant protection were invited to evaluate the outline model of the ontology. The experts believe that the ontology constructed in this study contains a comprehensive conceptual system for wheat sharp eyespot. It meets the evaluation criteria and the requirements for ontology construction, making it suitable for knowledge representation and the next stage of ontology application. The ontology model, refined based on expert evaluation, was visualized and expressed using Protégé (Figure 4).

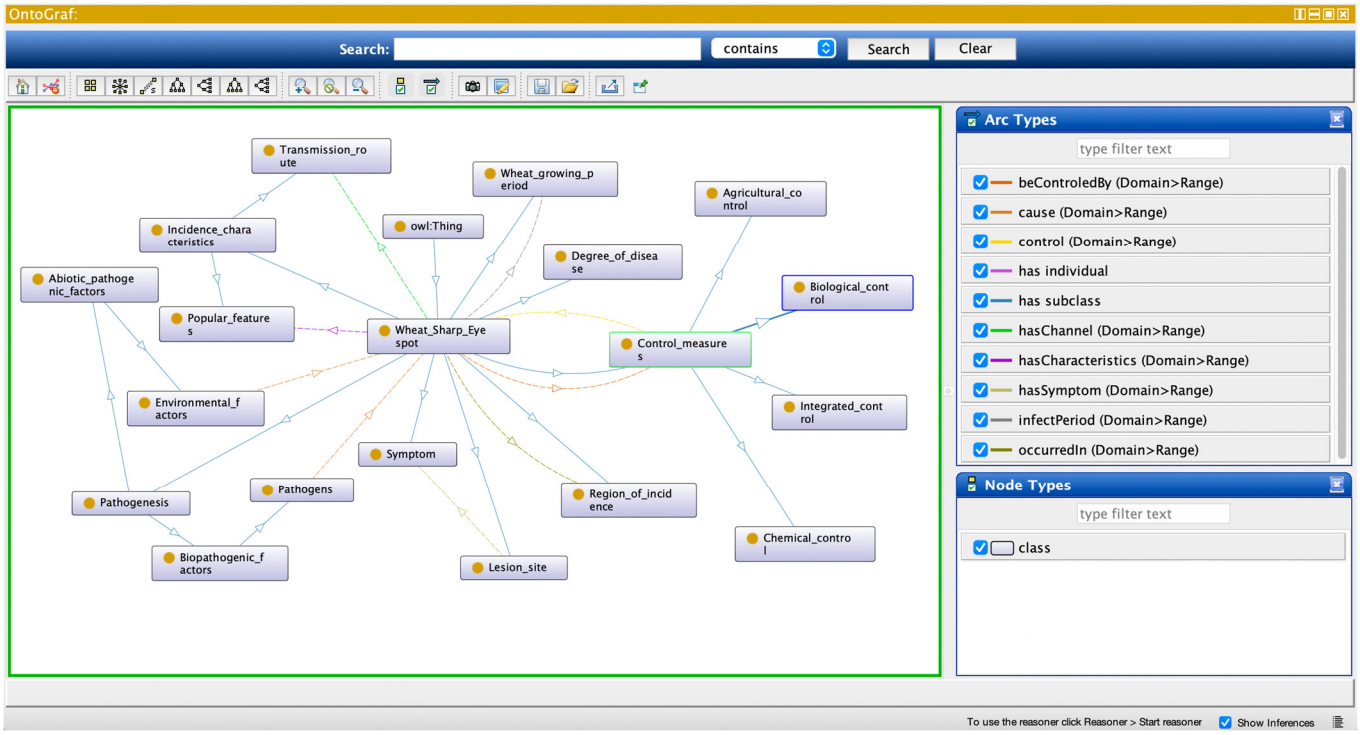

Figure 4. Ontology of wheat sharp eyespot control.

3.3. Automated Ontology-Based Knowledge Extraction

3.3.1. Model Training

In this study, a knowledge extraction model in the field of wheat sharp eyespot control was trained based on the ERNIE 3.0 knowledge enhancement pretraining model [31]. ERNIE 3.0 introduces large-scale knowledge graphs into pretraining models containing tens of billions of parameters. It proposes a parallel pretraining method that combines massive unsupervised text and large-scale knowledge graphs, known as Universal Knowledge–Text Prediction. Using a knowledge graph mining algorithm, it extracts 50 million knowledge graph triples and a 4TB large-scale corpus, which are simultaneously input into the pretraining model for joint mask training. This approach promotes information sharing between structured knowledge and unstructured text, significantly enhancing the model's memory and reasoning capabilities for knowledge.

The ERNIE 3.0 framework is divided into two layers (Figure 5). The first layer is the general semantic representation network, which learns basic and generalized knowledge from the data. The second layer is the task semantic representation network, which learns task-related knowledge based on the generic semantic representation. During the learning process, the task semantic representation network learns only the pretrained tasks of the corresponding category, while the generic semantic representation network learns all the pretrained tasks.

In this study, we selected the ERNIE 3.0 knowledge enhancement pretraining model to train a knowledge extraction model in the field of wheat sharp eyespot control. The main reasons for choosing ERNIE 3.0 include the following: (1) ERNIE 3.0 incorporates large-scale knowledge graphs, enabling the joint masked training of knowledge and text during the pretraining process. This promotes information sharing between structured knowledge and unstructured text, which is highly significant for knowledge integration and extraction in the agricultural field, especially for wheat sharp eyespot control; (2) ERNIE 3.0 is designed to significantly enhance the model's memory and reasoning abilities. This is crucial for application scenarios that require processing a large amount of complex agricultural knowledge and reasoning tasks, thereby better supporting the construction of agricultural expert systems; (3) the ERNIE 3.0 framework consists of two

layers: the general semantic representation network and the task semantic representation network. These layers can learn basic and task-related knowledge, respectively. This structure improves the model's adaptability and generalization capabilities across different tasks, making it not only suitable for wheat sharp eyespot control but also extendable to other agricultural disease areas.

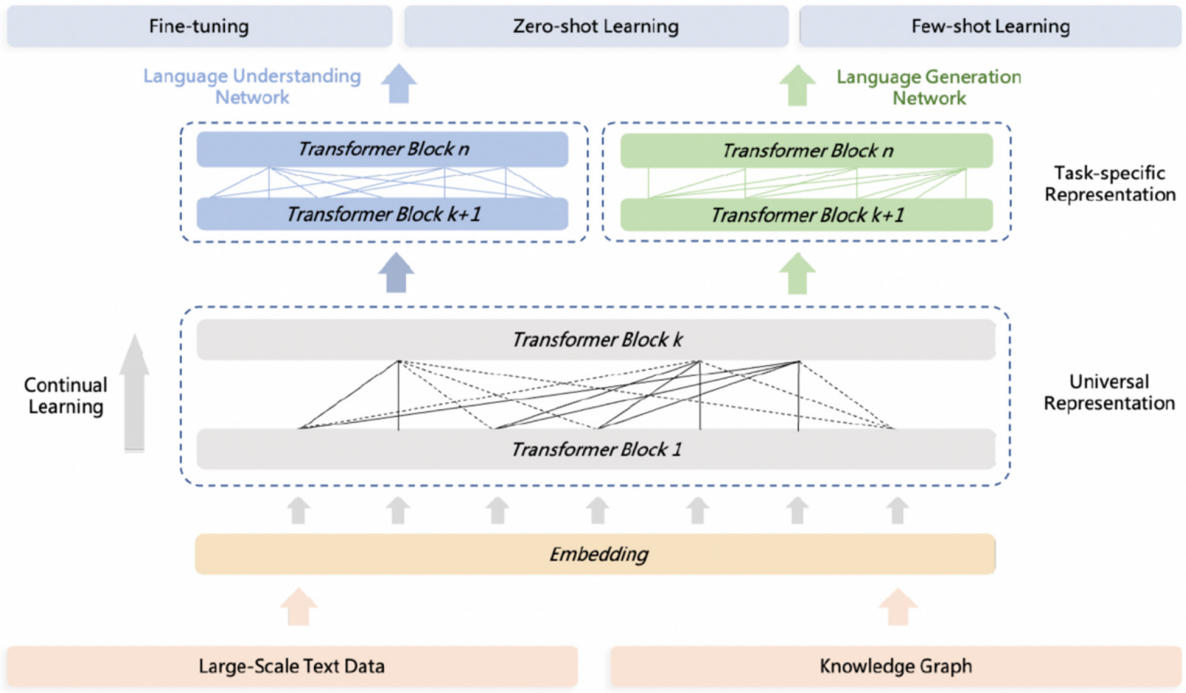

Figure 5. ERNIE 3.0 model framework.

In this study, we used the data annotation platform Doccano for data annotation and divided the data into a training set, validation set, and test set according to the ratio of 8:1:1. The appropriate construction of negative examples can enhance the model's effectiveness. The ratio of positive and negative samples in the training set was 5:1 in order to ensure the accuracy of the evaluation index. The validation and test sets were constructed with full negative examples by default. Table 5 presents the descriptive statistics of the datasets.

Table 5. Descriptive statistics of datasets.

Dataset	Total Samples	Positive Samples	Negative Samples
Training Set	2176	1813	363
Validation Set	272	272	0
Test Set	272	272	0

The learning rate was dynamically adjusted by using the strategy of cosine annealing [32] as follows:

$$\eta_t = \frac{1}{2}(\eta_{max} - \eta_{min})\left(1 + \cos\left(\frac{T_{cur}}{T_i}\pi\right)\right) \quad (2)$$

The initial value of η_{max} is the learning rate, T_{cur} is the current number of training rounds in the training process of SGDR (restart training SGD), and T_i is the number of epochs between two restarts of SGDR. When $T_{cur} = T_i$, set $\eta_t = \eta_{min}$. When $T_{cur} = 0$ after restart, set $\eta_t = \eta_{max}$.

To accurately evaluate the performance advantages and disadvantages of the model, we used three fundamental evaluation metrics in the field of entity–relationship extraction: precision, recall, and *F1* score. These metrics were employed to assess the model's

performance comprehensively. The calculation method of each evaluation index was as follows:

- Recall: This metric determines the proportion of true facts that have actually been denoted as true by the model. Considering TP and FN as the number of true facts correctly and incorrectly classified, respectively, the recall can be obtained as follows:

$$R = \frac{TP}{TP + FN} \tag{3}$$

- Average Precision: This metric weights the precision and recall increment of the model at different threshold values n. It provides an overall measurement of the model's classification performance while penalizing biased predictions. It is calculated as follows:

$$AP = \sum_{n} (R_n - R_{n-1}) P_n \tag{4}$$

where R refers to the recall value, and P represents the precision, computed as follows:

$$P = \frac{TP}{TP + FP} \tag{5}$$

where FP denotes the number of unfeasible facts predicted as true.

- *F1* Score: This metric is the harmonic mean between precision and recall and serves as an indicator of the model's accuracy. It can be calculated using the following equation:

$$F1 = \frac{2(P \times R)}{(P + R)} \tag{6}$$

The *F1* value is calculated between the predicted ternary and the gold ternary, and the prediction is considered correct when the predicted [pred_head, pred_rel, pred_tail] is exactly the same as the gold [head, rel, tail].

The model training process is shown in Figure 6. Compared to the baseline model UIE, the accuracy of our model increased by 13.89%, its recall increased by 22.27%, and the *F1* score increased by 17.86%. Table 6 shows the performance metrics of our model compared to the baseline model UIE.

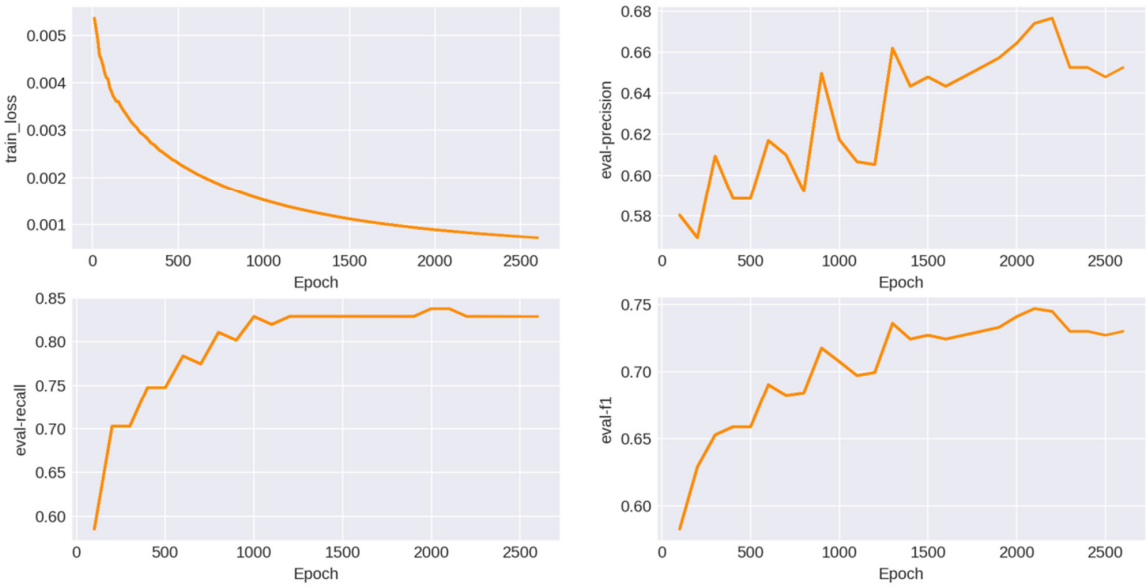

Figure 6. Model training process.

Table 6. Model performance metrics.

Metric	Our Model	Baseline Model UIE	Improvement
Precision/%	87.04	73.15	+13.89
Recall/%	95.92	73.65	+22.27
F1 Score/%	91.26	73.40	+17.86

3.3.2. Model-Based Reasoning

The wheat sharp eyespot domain knowledge extraction model was designed for document-level extraction tasks. For these tasks, a sliding window was used to combine contextual sentences and divide them into smaller paragraphs. This approach helps maintain the contextual coherence of the extracted knowledge fragments and reduces ambiguity. The sliding window covers different parts of the document to ensure that the extracted knowledge is both representative and diverse. Additionally, this method allows the model to process the entire document step by step without experiencing performance degradation or memory shortages due to the document's length. The specific inference algorithm is illustrated in Figure 7.

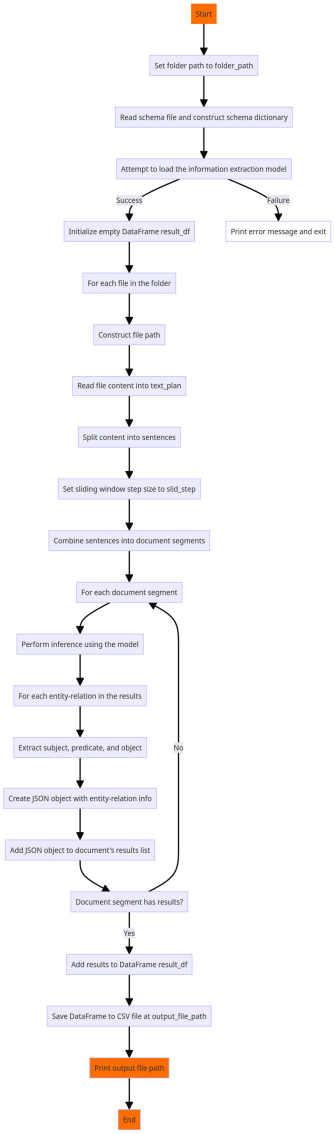

Figure 7. Inference workflow.

4. Discussion

The data sources for the ontology in the field of wheat sharp eyespot control, as constructed in this study, were primarily derived from the scientific and technical literature. In the future, the ontology can be extended by expanding the data sources. For example, credible resources on the Internet, patent databases, reports, and websites of agricultural research organizations can be utilized to collect the latest progress and new findings on wheat sharp eyespot control. During the actual application of the ontology, it needs to be constantly updated and improved. As research progresses, new concepts and relationships may emerge. When collecting new scientific and technological research and data, it is essential to incorporate this new information into the ontology in a timely manner to ensure its completeness. The emergence of large language models, such as the GPT, provides brand new possibilities for the extension of the ontology and the extraction of knowledge. Large language models can extract important information about wheat sharp eyespot control from extensive scientific and technical research, patent databases, and reports from other agricultural research institutions. They can integrate domain terminology and new concepts, facilitating the automated updating, optimization, and extension of knowledge. There have been related studies in the fields of biology and medicine, such as SPIRES, which involves a knowledge extraction algorithm based on a large language model [33] and genome aggregation using the GPT model as a complement to standard enrichment analysis [34]. The application of large language modeling in knowledge updating in the field of wheat sharp eyespot control deserves further exploration and practice.

5. Conclusions

In this study, we proposed an approach for ontology construction in the field of wheat sharp eyespot and developed an ontology-based knowledge extraction model that can efficiently extract relevant knowledge from text. This approach has significant advantages in integrating and organizing dispersed information and provides strong support for applications such as intelligent Q&A and intelligent recommender systems in the agricultural domain, thus facilitating knowledge discovery and sharing. Although the model performs well in the field of wheat sharp eyespot control, its effectiveness is dependent on the quality of the input data, and the scope of applicability needs to be further validated to confirm its applicability to different crops and agricultural diseases. Future work will focus on the following areas: expansion of the dataset to further collect more diverse and extensive datasets to improve the robustness and accuracy of the knowledge extraction model; the validation of cross-domain applicability, with a focus on adapting and validating the applicability of our knowledge extraction framework to other agricultural domains; and integration of the model with advanced decision support systems to better model the experts' decision-making process, help farmers and agricultural practitioners cope with wheat sharp eyespot more effectively, and promote sustainable agricultural development.

Author Contributions: K.L.: data curation, formal analysis, investigation, methodology, writing—original draft, writing—review and editing. Y.C.: funding acquisition, project administration, supervision, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the National Key Research and Development Program of China (2022YFF0711902-01), the Central Public-Interest Scientific Institution Basal Research Fund (Y2024XK07), the Fundamental Research Funds for AII-CAAS (JBYWAII202331), and Beijing Municipal Innovation Team Building Project for the Modern Agricultural Industry Technology System (BAIC10-2023-E10).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article. Further inquiries can be directed to 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 a potential conflict of interest.

References

1. Boerema, G.H.; Verhoeven, A.A. Check-list for scientific names of common parasitic fungi. series 2b: Fungi on field crops: Cereals and grasses. *Neth. J. Plant Pathol.* **1977**, *83*, 165–204. [CrossRef]
2. Jia, T.; Wu, G.; Liu, C. Current status of research on root rot diseases of wheat in China and countermeasures for their prevention and control. *Chin. Agric. Sci.* **1995**, *3*, 41–48.
3. Pan, H. Occurrence of Wheat Sharp Eyespot and its control measures. *Henan Agric.* **2016**, *7*, 39.
4. Yao, L.; Wang, Q.; Fu, X.; Mei, R. Screening and characterization of Bacillus cereus against Wheat Sharp Eyespot. *China Biol. Control.* **2008**, *24*, 53–57.
5. Zheng, Y.; Zhu, D.; Wu, H.; Peng, X. A review of the knowledge graph Q&A domain. *Comput. Syst. Appl.* **2022**, *31*, 1–13.
6. Li, J. Research on the Construction Method and Application of Domain Ontology. Ph.D. Thesis, Chinese Academy of Agricultural Sciences, Beijing, China, 2009.
7. Consortium, T.P.O. The Plant Ontology™ Consortium and Plant Ontologies. *Comp. Funct. Genom.* **2002**, *3*, 137–142. [CrossRef] [PubMed]
8. Chang, C. Food and Agriculture Organization of the United Nations AOS Project. *J. Agric. Libr. Inf.* **2003**, *2*, 14–15+24.
9. Dai, C.-P.; Huang, Y.-D.; Qian, P.; Wang, R.J.; Dong, W.; Huang, Q. Research on the construction of rice pest and weed ontology. *Guangdong Agric. Sci.* **2011**, *38*, 191–194.
10. Yu, H.; Shen, J.; Bi, C.; Liang, J.; Chen, H. Intelligent diagnosis system for rice pests and diseases based on knowledge graph. *J. South China Agric. Univ.* **2021**, *42*, 105–116.
11. Qi, H.; Guan, Y.; Liu, Y. An ontology learning study of corn pests and diseases for Chinese text. *Comput. Eng. Appl.* **2011**, *47*, 206–209.
12. Zhang, L.; Duan, Q.; Li, D. Research on ontology construction technology for diagnosis and treatment of corn pests and diseases. *Agric. Mech. Res.* **2012**, *34*, 41–45.
13. Liu, G.; Yang, Q.; Liu, Q. Ontology construction and visualization of agricultural science datasets--Taking the field of "cotton disease control" as an example. *J. Intell.* **2022**, *41*, 143–149+175.
14. Ren, N.; Sun, Y.; Bao, T.; Guo, T. Research on ontology construction method in agricultural domain--Taking tomato pests and diseases as an example. *Intell. Explor.* **2021**, *7*, 51–57.
15. Deepa, R.; Vigneshwari, S. An effective automated ontology construction based on the agriculture domain. *ETRI J.* **2022**, *44*, 573–587. [CrossRef]
16. Xu, D.; Lu, W.; Xu, R.; Zhnag, H.; Jiang, Y.; You, L.; Feng, Z. A decision-making method for precision fertilization of rice based on spatio-temporal multimodal knowledge mapping in agriculture. *J. Huazhong Agric. Univ.* **2023**, *42*, 281–292.
17. Zeng, D.; Liu, K.; Lai, S.; Zhou, G.; Zhao, J. Relation classification via convolutional deep neural network. In Proceedings of the COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, Dublin, Ireland, 23–29 August 2014.
18. Makoto, M.; Bansal, M. End-to-End Re-lation Extraction using LSTMs on Sequences and Tree Structures. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016.
19. Dong, L.; Lapata, M. Language to Logical Form with Neural Attention. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016.
20. Zheng, S.; Wang, F.; Bao, H.; Hao, Y.; Zhou, P.; Xu, B. Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, BC, Canada, 30 July–4 August 2017.
21. Luan, Y.; He, L.; Ostendorf, M.; Hajishirzi, H. Multi-Task Identification of Entities, Relations, and Coreference for Scien-tific Knowledge Graph Construction. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018.
22. Wang, J.; Lu, W. Two are better than one:Joint entity and relation extraction with table sequence encoders. *arXiv* **2020**, arXiv:2010.03851.
23. Sui, D.; Chen, Y.; Liu, K.; Zhao, J. Joint Entity and Relation Extraction with Set Prediction Networks. *arXiv* **2020**, arXiv:2011.01675. [CrossRef] [PubMed]
24. Moreno, A.; Isern, D.; Lpez Fuentes, A.C. Ontology-Based Information Extraction of Regulatory Networks from Scientific Articles with Case Studies for Escherichia Coli. *Expert Syst. Appl.* **2013**, *40*, 3266–3281. [CrossRef]
25. Li, C.X.; Su, Y.R.; Wang, R.J.; Wei, Y.Y.; Huang, H. Structured AJAX Data Extraction Based on Agricultural Ontology. *J. Integr. Agric.* **2012**, *11*, 784–791. [CrossRef]
26. Wimalasuriya, D.C.; Dou, D. Using Multiple Ontologies in Information Extraction. In Proceedings of the 18th ACM Conference on Information and Knowledge Management, Hong Kong, China, 2–6 November 2009; pp. 235–244. [CrossRef]
27. Zhang, L. Research on Ontology Construction Methods for Agricultural Domain Based on Narrative Lists and Literature Databases. Master's Thesis, Chinese Academy of Agricultural Sciences, Beijing, China, 2011.

28. Yang, J. Ontology-Based Knowledge Modeling and Reasoning for Citrus Pests and Diseases. Master's Thesis, Central China Normal University, Wuhan, China, 2014.
29. Li, J.; Meng, L. A Comparative Study of Methodological Systems for Constructing Knowledge Ontologies. *Mod. Libr. Intell. Technol.* **2004**, *7*, 17–22.
30. Song, J. Research on Agricultural Knowledge Mapping Construction Based on Knowledge Distillation. Master's Thesis, Harbin Institute of Technology, Harbin, China, 2022.
31. Wang, S.; Sun, Y.; Xiang, Y.; Wu, Z.; Ding, S.; Gong, W.; Feng, S.; Shang, J.; Zhao, Y.; Pang, C.; et al. Ernie 3.0 titan: Exploring larger-scale knowledge enhanced pre-training for language understanding and generation. *arXiv* **2021**, arXiv:2112.12731.
32. Loshchilov, I.; Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv* **2016**, arXiv:1608.03983.
33. Caufield, J.H.; Hegde, H.; Emonet, V.; Harris, N.L.; Joachimiak, M.P.; Matentzoglu, N.; Kim, H.; Moxon, S.; Reese, J.T.; Haendel, M.A.; et al. Structured prompt interrogation and recursive extraction of semantics (SPIRES): A method for populating knowledge bases using zero-shot learning. *arXiv* **2023**, arXiv:2304.02711. [CrossRef]
34. Joachimiak, M.P.; Caufield, J.H.; Harris, N.L.; Kim, H.; Mungall, C.J. Gene Set Summarization using Large Language Models. *arXiv* **2023**, arXiv:2305.13338v2.