Communication Papers of the 18th Conference on Computer Science and Intelligence Systems pp. 141–146
DOI: 10.15439/2023F6100
ISSN 2300-5963 ACSIS, Vol. 37

Automatic Construction of Knowledge Graph of Tea Diseases and Pests

1st Qiang Huang
0000-0002-2552-2687
College of Information engineering
Sichuan Agricultural University
Yaan City, China
13499@sicau.edu.cn

2nd YouZhi Tao
0009-0009-5429-4888
College of Information engineering
Sichuan Agricultural University
Yaan City, China
taoyouzhi@stu.sicau.edu.cn

3nd Shitao Ding
0009-0000-9886-5077
College of Information engineering
Sichuan Agricultural University
Yaan City, China
2020319014@stu.sicau.edu.cn

4th Yongbo Liu
0009-0000-4009-680X
Agricultural Information and Rural Economy Institute
Sichuan Academy of Agricultural Sciences
Chengdu City, China
182382602@qq.com

corresponding: Francesco Marinello
0000-0002-3283-5665
University of Padova
Padua City, Italy
francesco.marinello@unipd.it

***Abstract*—Tea production involves several stages, usually pests and diseases can negatively impact the quality of tea and reduce the harvest, limiting the industry's development. Therefore, it is important to unify the knowledge on tea pests and diseases. Unfortunately, the current knowledge graph construction for tea pests and diseases relies mainly on semi-automated and manual methods, resulting in inefficiency and failing to meet production demands. This research combines three model of Bidirectional Encoder Representation from Transformers, Bi-directional Long Short-Term Memory, Conditional Random Fields for joint extraction of data. The model abbreviated as BERT-BiLSTM-CRF, using the model can automatically generates the triplets, and then store them in the Neo4j database. The study shows that this model has improved accuracy compared to previous methods, and provides effective support for scientific management and production services of tea pests and diseases. The research offers a reference for quickly constructing knowledge graphs in the crop domain.**

***Index Terms*—tea pests and diseases, knowledge graph, BERT-BiLSTM-CRF, Neo4j**

I. Introduction

THE knowledge graph is a structured semantic knowledge base, which is essentially a semantic network that describes the relationships between entities, and can now be used to refer to various large-scale knowledge bases. The knowledge graph represents the relationships between entities in the form of an entity-relationship-entity triplet [1-2] and has been widely used in the fields of research, internet and artificial intelligence [3-4]. With the continuous development of artificial intelligence, machine learning, big data and other disciplines, knowledge graph has achieved better results in domain knowledge management, and the construction of knowledge graph in specific areas of agriculture has gradually become the focus of research by researchers at home and abroad. Yongbo Liu constructed a tea knowledge graph based on the BERT-WWM and attention mechanism approach [5]. Haussmann constructed a knowledge graph of agricultural information [6], which enables users to select the available agricultural products to make food. Dandan Wang combined 2 methods, bottom-up and top-down, to construct a knowledge graph of rice [7], Xu Xin used Neo4j and NLP technology to construct a knowledge graph of wheat varieties [8], which solved the problem of high knowledge repetition rate and unclear knowledge association in variety data. In the process of tea production and marketing, we will face several aspects such as planting, management and processing, each of which requires scientific technical guidance. In actual production, tea yield reduction caused by pests and diseases is generally 15%-20%, which can lead to no tea harvesting in serious cases [9]. Pests and diseases are important factors limiting the development of the tea industry. However, the currently existing knowledge graph in the field of tea pests and diseases are mainly constructed in a semi-automatic and manual way, with low construction efficiency, which cannot meet the actual production needs.

In this paper, we constructed a domain text dataset based on ME+R+BIESO annotation, and used the BERT-BiLSTM-CRF model for joint triplet extraction of entities and relationships from unstructured data to realize the automated construction of knowledge graph, which provides a reference basis for the rapid construction of knowledge graph in crop domains.

II. Tea Pest And Disease Knowledge Graph Construction Process

The knowledge graph can be divided into general domain knowledge graph and vertical domain knowledge graph according to different application directions [10-11]. General domain knowledge graph has the characteristics of being oriented to the whole domain, having a wide range of audiences and involving shallow industry knowledge, and are mostly used in business scenarios such as Internet search engine and content recommendation, e.g., Google search en-

Funded by: Sichuan Provincial Financial Independent Innovation Special Plan Project - Application Research of Spatiotemporal Big Data Analysis in Agricultural Production Services (2022ZZCX034), China

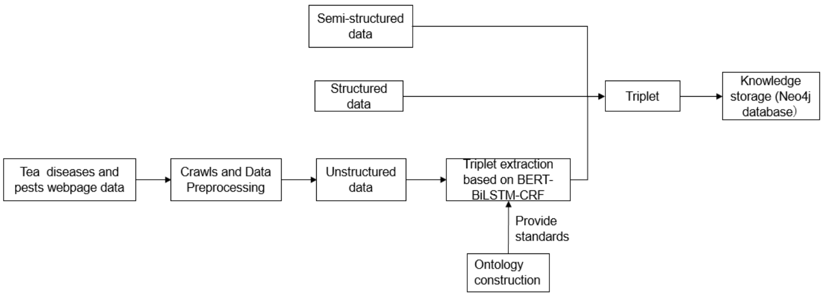
Semi-structured data
Structured data
Triplet
Knowledge storage (Neo4j database)
Tea diseases and pests webpage data
Crawls and Data Preprocessing
Unstructured data
Triplet extraction based on BERT-BiLSTM-CRF
Provide standards
Ontology construction

Figure 1. Process for constructing a knowledge graph of tea pests and diseases

gine, FreeBase [12], DBpedia [13], etc. In this study, the knowledge graph of tea pests and diseases belongs to the vertical domain, and the top-down approach is used to build the graph ontology. This approach requires defining the ontology and data schema first, and then populating the entities and their relationships into the knowledge graph. As shown in Figure 1, it specifically includes the following five stages: (1) Data acquisition and processing. Data is obtained from a variety of sources, which include structured, semi-structured and unstructured data. Structured data can be obtained from third party databases, while semi-structured data typically includes HTML web pages and JSON data. In this paper, unstructured data is obtained from the tea pest knowledge website and data cleaning and pre-processing operations are carried out to obtain the raw text data. (2) Ontology construction. Construct tea pest and disease ontology based on domain corpus, define classes, relations and attributes, and set corresponding constraints to clarify the boundary of knowledge extraction; (3) Data annotation. Use Brat data annotation tool to annotate the text set and obtain the relevant training set and test set after processing by Python code; (4) Knowledge extraction. The BERT-BiLSTM-CRF model is used for training, and the trained model is used to do triplet extraction of the data set. (5) Knowledge storage. The extracted tea pests and diseases triplet data are stored in the Neo4j graph database [14] and visualized.

A. Data Acquisition and Pre-processing

The data related to tea pests and diseases in this paper are mainly from the website China Crop Germplasm Information Network, which comes from the Institute of Crop Science, Chinese Academy of Agricultural Sciences. The page contains information on pests and diseases of various crops, and the data on tea pests and diseases mainly contains information on symptoms, alias, pathogen categories and other attributes. The website contains data of 71 tea pests and 21 tea diseases. The Scrapy crawler framework is used for data crawling, and the data pre-processing is combined with rules and manual review to obtain a noise-free plain text corpus.

B. Ontology Construction

The architecture of the knowledge graph is generally divided into two layers: the schema layer and the data layer. The schema layer is the core of the knowledge graph structure and is built on top of the data layer. Designing the schema layer of tea pest and disease knowledge graph before data extraction is beneficial to reduce data redundancy. According to the data characteristics of tea pests and diseases, the tea disease knowledge graph concept and tea pest knowledge graph concept are designed respectively, as shown in Figure 2.

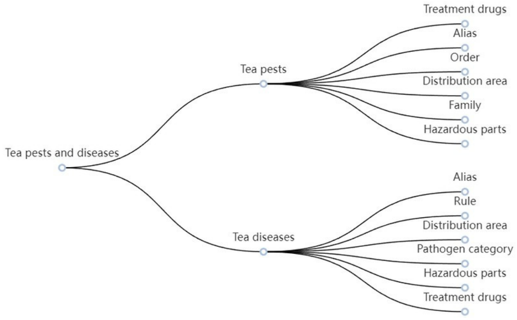

Figure 2. Conceptual level of tea pest and disease knowledge graph

C. Data Annotation

ME+R+BIESO annotation method is used to annotate the main entity and the relationship between the main entity and other entities. First, the main entity is labeled as “ME”, and when there is a relationship between an entity and the main entity in a piece of data, the entity Ei is labeled as relationship Ri. The information of each character in the entity is indicated by using the Begin-Inside-End-Single-Other, BIESO) flag to indicate the information of each character in this entity. When the complete set of BE, BIE or S of the main entity ME and a certain relation Ri is matched, the main entity and Ei corresponding to this tag set are taken out and the (ME, Ri, Ei) triplet is formed by data parsing.

Take the data of tea pest “Artaxa flava” as an example, as shown in Figure 3. First of all, “Artaxa flava” is labeled as ME (Main Entity), and “yellow poisonous moth” is an alias of “Artaxa flava”, so “yellow poisonous moth” is labeled as alias. After the data labeling task is completed, the generated a file and the original txt text file are used to label each character in the text with a corresponding label using Python code, and other irrelevant characters are labeled as “O”. When matching the main entity ME and the set of BIE or BE tags with the relationship “Alias”, the mapping of tags can generate a triplet (Artaxa flava, Alias, yellow poisonous moth).

The ME+R+BIESO annotation method focuses on the annotation of the relationship type Ri between the main entity and other entities, without focusing on the entity type to which the entity itself belongs. This method only annotates and extracts on a predefined set of relationships to reduce redundancy and error propagation of irrelevant entity pairs. For the case of overlapping relationships between the main entity ME and multiple entities Ei, multiple corresponding triplets can be obtained by label matching and mapping.

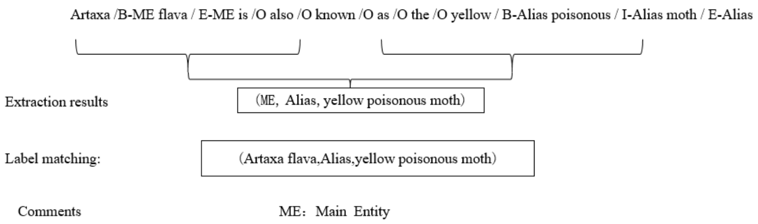
Artaxa /B-ME flava / E-ME is /O also /O known /O as /O the /O yellow / B-Alias poisonous / I-Alias moth / E-Alias
Extraction results
（ME，Alias, yellow poisonous moth）
Label matching:
（Artaxa flava,Alias,yellow poisonous moth）
Comments
ME：Main Entity

Figure 3. Example of ME+R+BIESO labeling method

Compared with the traditional entity relationship extraction methods, the ME+R+BIESO method can synchronize the labeling of bodies and relationships, which reduces the labeling cost and improves the efficiency.

D. BERT-BiLSTM-CRF Model

For the upstream task of entity recognition, traditional corpus learning models such as Word2Vec [15], Glove [16] and other single-layer neural networks cannot characterize the multi-sense of words well in Chinese language environment, so this study chooses the Bidirectional Encoder Representations from Transformers as the linguistic pre-processing model for graph construction. Representations from Transformers as the language preprocessing model for graph construction, in order to obtain high-quality word vectors for entity extraction and classification of downstream tasks. In 2015, the BiLSTM-CRF [17] model proposed by Baidu Research Institute was used for named entity recognition.

A BiLSTM-CRF end-to-end model based on BERT word embedding is used to train and predict tags based on the ME+R+BIESO annotation model. The model consists of three components: namely the BERT layer, the bi-directional LSTM layer and the CRF layer [18], and the overall framework of the model is shown in Figure 4. Firstly, the previously annotated corpus is encoded by the BERT pre-training model to extract the tea pest text corpus features and generate word vectors corresponding to words based on the contextual features of the current words. The key part of the BERT pre training model is in the Transformer layer. The core of the Transformer layer is to calculate the correlation between words through the self-attention function Attention, in order to allocate the weight of words [19].

$$\text{Attention }(Q,K,V) = \text{softmax}\left(\frac{Qk^t}{\sqrt{d_k}}\right)V \tag{1}$$

In the equation, headi represents single headed Attention, and MultiHead represents multiple

head attention, W is the weight matrix, through multiple different linear variables change the projection of Q, K and V, and then use the concatenation function concat to concatenate the results of the self-attention mechanism by multiplying them by weights, and calculate the position information of different spatial dimensions.

$$\text{head}_i = \text{Attention }\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{2}$$

$$\begin{gathered}\text{MultiHead(Q, K, V)} = \\ \text{Concat}(head_1\ , head_2\ \ldots, head_n\ )\ W^0\end{gathered} \tag{3}$$

In the formula, Q, K, and V are all word vector matrices, and dk represents the input dimension,

WiQ, WiK, WiV represent the weight matrix, and W0 represents the additional weight matrix.

Obtain the word vector corresponding to the input sequence through the BERT model, and then input the word vector into the BiLSTM module for bidirectional encoding. The BiLSTM model overcomes the dependency limitations

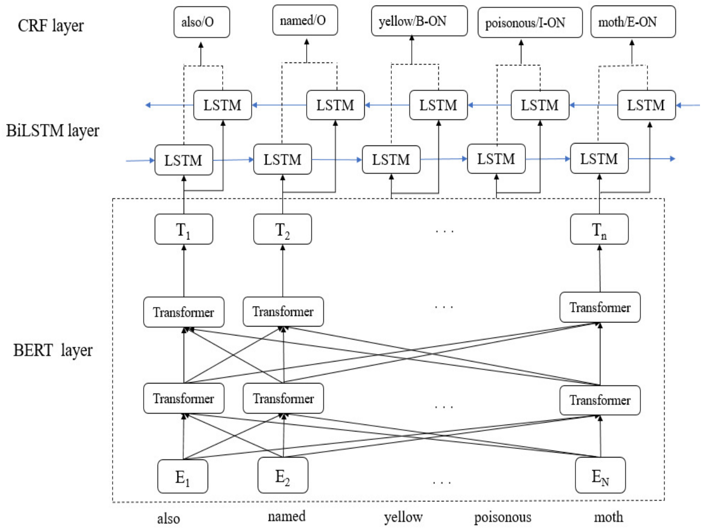

Figure 4. Overall framework of BERT-BiLSTM-CRF model

of traditional machine learning, and bidirectional encoding allows contextual information to be read into the model to achieve effective prediction of tag sequences, and the model outputs a predicted score value for each tag. Finally, the output of the BiLSTM model is decoded using the CRF model to obtain the final predicted annotated sequence. Compared with general deep learning named entity recognition models, the most important feature of this model is the incorporation of a BERT pre-training model, which does not require pre-training of word vectors, and the rich word-level features, syntactic structure features and semantic features of the sequences can be extracted by directly feeding the sequences into the BERT model.

E. Triplet Extraction

The trained model was saved and automated triplet extraction of unstructured text was performed. The automated extraction process is shown in Figure 5, using data from pest-related websites and using the Scrapy crawler framework to automatically obtain unstructured text. In the web data, a page is described for the same tea pest content. To improve the accuracy of the triplet extraction, the main entity, the tea pest name entity, can be identified using a split word method. Now only the corresponding relationships and tail entities need to be predicted using the model. The corresponding label for each word is obtained from the saved model predictions, and the predicted labels "B-Ri", "I-Ri" and "E-Ri" are then used according to the predicted labels "B-Ri", "I-Ri" and "E-Ri" are combined in the order of "BIE" or "BE" to form the corresponding tail entity Ei, which forms a ternary data set of the form (ME, Ri, Ei). Combined with the mapping of the relationship type corresponding to the relationship label R and the custom entity label dictionary, it is converted into the final (main entity, relationship, tail entity) form to complete the automated extraction of the triplet.

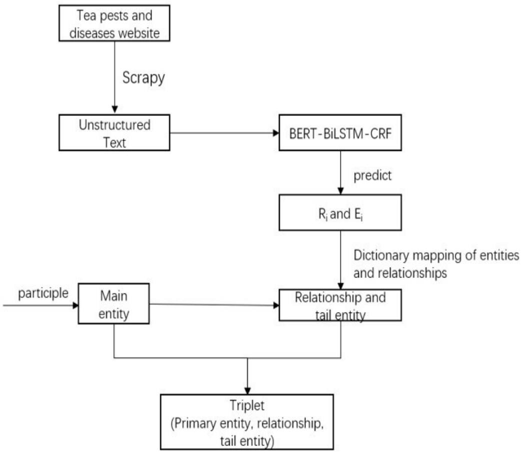
Tea pests and diseases website
Scrapy
Unstructured Text
BERT-BiLSTM-CRF
predict
R_i and E_i
Dictionary mapping of entities and relationships
participle
Main entity
Relationship and tail entity
Triplet (Primary entity, relationship, tail entity)

Figure 5 Automatic extraction of triplets

F. Knowledge Storage

Neo4j is a popular graph database that can store entities, attributes, and relationships in the knowledge graph using graphical representation. This storage mode makes visualization and query of knowledge graph very convenient. Use the Neo4j graph database for knowledge storage, as shown in Figure 6, which is a visual knowledge graph of tea pests "Aleuro10bus marlatti Quaintance", "Aonidiella aurantia" and "Artaxa flav". The data statistics of the triplets are shown in Table I.

III. Experimental Results And Analysis

The experimental environment is Python 3.8 and Pytorch 1.8, the model uses precision, recall and F1 values as evaluation metrics.

A. Comparison of different models

In order to validate the BERT-BiLSTM-CRF model of superiority, four groups of experiments were set up in this paper. The LSTM, LSTM-CRF and BiLSTM-CRF models were chosen as control experiments respectively; the results of the model comparison experiments are shown in Figure 7.

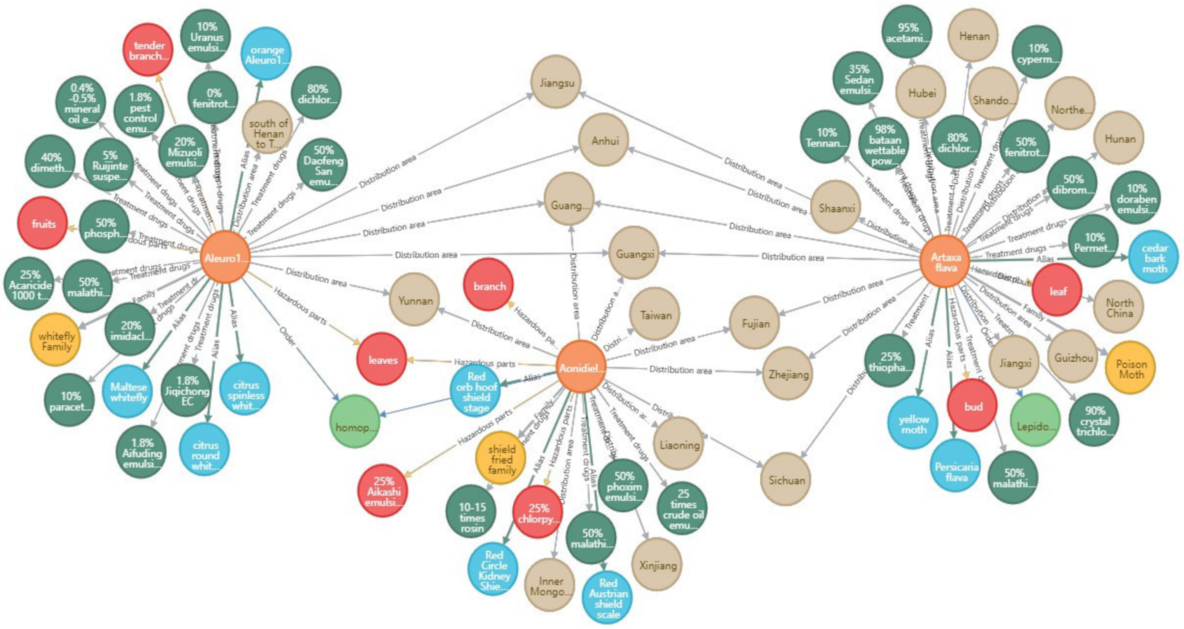

Figure 6. Example of tea pest data visualization

Table I.
Triplet data statistics

Name	Quantity	Meaning
Alias	200	Alias information for tea disease or pest entities
Order	70	Biological classification of tea pest entities "Order"
Family	71	Biological classification of tea pest entities "Family"
Treatment drugs	572	Information on drugs for the control of tea diseases or pest entities
Distribution area	662	Information on the regional distribution of tea disease or pest entities
Hazardous parts	176	Information on the damage sites of tea diseases or pest entities
Pathogen category	20	Information on the pathogenic categories of tea disease entities
Rule	36	Information on the occurrence pattern of tea disease entities
Total	1807	Total number of triplets

Compared with BiLSTM-CRF and LSTM-CRF, the BERT-BiLSTM-CRF model improves the accuracy by 1.76%~5.48%, the recall by 4.25%~8.61%, the F1 score by 3%~7.05%, and the F1 score by 90.53%. The BERT-BiLSTM-CRF model improved the F1 score by 3% after adding the BERT pre-trained language model to the BiLSTM-CRF, indicating that BERT can assist in improving the model's semantic representation of the text and capture the interrelated entity relationships in the tea pest text to a greater extent, thus optimizing the entity effect of the relationship extraction "task".

B. Prediction results of different relationships

The prediction results of the BERT-BiLSTM-CRF model for the relationship between the main entity and each entity are shown in Figure 6, which shows that the overall effect is good, with an F1 score of 90.53%. The "Order", "Family", "Alias", "Distribution area", "Treatment drugs" and "Pathogen category" of tea pests and diseases are shown in Figure 8, The six types of relationships, "Order", "Family", "Alias", "Distribution area", "Treatment drugs" and "Pathogen category", were identified well. In particular, the prediction accuracy of "Order" and "Family" was close to 100%, because the data characteristics of these two types of relationships were very obvious. The BERT-BiLSTM-CRF model can effectively learn the textual information. However, the prediction results of the relationship between "Hazardous parts" and "Rule" were significantly lower than the average. By analyzing the text of the corresponding corpus and the final prediction results of the relationship between "Hazardous parts" and "Rule", we can see that the damage sites of different pests and diseases in this paper are different, such as "leaf" , "stems", "branches", "tea tree root system" and other words are describing the damage site; the same as "relative humidity 85%-87%", "poor ventilation and light penetration", "temperature 25-28℃", "high temperature and low humidity", etc. are all describing the occurrence pattern of the disease. The inconsistency of description methods makes it difficult for the model to fully learn the characteristics of the damage site, which makes the recall rate of "Hazardous parts" and "Rule" is low.

In this paper, the F1 values of the named entity recognition model basically reached more than 90% for entities other than "Hazardous parts" and "Rule". In summary, the BERT-BiLSTM-CRF model in this paper has a relatively good recognition effect in the named entity recognition task of tea pests and diseases.

IV. Conclusion

The BERT-BiLSTM-CRF model used in this paper extracts the triplet data of tea pests and diseases and automates the construction to generate the knowledge graph of tea pests and diseases. The experimental results show that the accuracy value reaches 90.10% and the recall rate

is 90.53%. It provides effective support for the scientific management and production services of tea pests and diseases, and the study also provides a reference basis for the rapid construction of knowledge graph in crop fields.

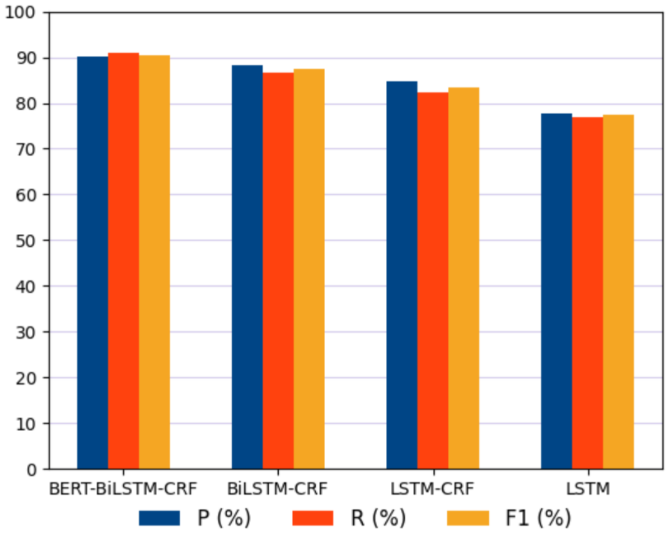

Figure 7. Entity extraction model performance comparison

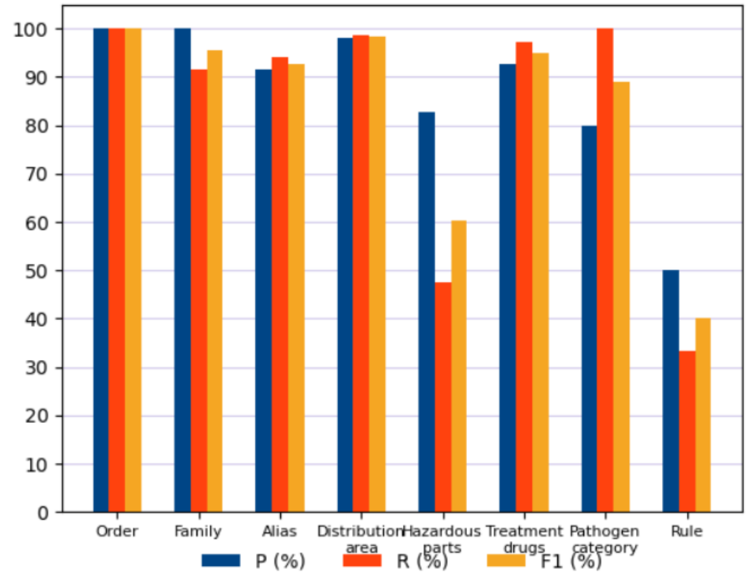

Figure 8. Prediction results of each relationship in the BERT-BiLSTM-CRF model

References

[1] Xu Zenglin, Sheng Yongpan, He Lirong, et al. "A review of knowledge graph techniques," Journal of University of Electronic Science and Technology of China, 2016, 45(4):18.

[2] Paulheim H, "Knowledge graph refinement: A survey of approaches and evaluation methods," Semantic web, 2017, 8(3): 489-508, to be published.

[3] Pujara J, Hui M, Getoor L, "Large-Scale Knowledge Graph Identification using PSL" Aaai Fall Symposium, 2013, to be published.

[4] Zhang Qingling, Li Xianzheng, Li Hangyu, et al. "Application of knowledge graph in agriculture," Electronic Technology & Software Engineering, 2019(7):3.

[5] Liu YB, Huang Q, Gao WB, et al. "Construction of tea knowledge graph by integrating BERT-WWM and attention mechanism," Southwest Journal of Agriculture,2022,35(12):2912-2921.

[6] Haussmann S, Seneviratne O, Chen Y. "Food KG: a semantics-driven knowledge graph for food recommendation," International Semantic Web Conference. Springer, Cham, 2019: 146-162, to be published.

[7] Wang Dandan, "Research and application of knowledge graph construction method for Ningxia rice," Northern University for Nationalities, 2020.

[8] Xu Xin, Yue Jinzhao, Zhao Jinpeng, et al. "Research on the construction and visualization of knowledge graph of wheat varieties," Computer System Applications,2021,30(06):286-292.

[9] Tan Rongrong, Liu Mingyan, Gong Ziming, et al. "Analysis of the types and occurrence patterns of major pests and diseases in tea areas of Hubei Province," Tea Newsletter, 2013, 40(04):36-38.

[10] MA Mohamed, Pillutla S, "Cloud computing: a collaborative green platform for the knowledge society," Vine, 2014, 44(3): 357-374, to be published.

[11] Hu Fanghuai, "Research on Chinese knowledge graph construction method based on multiple data sources," Shanghai: East China University of Science and Technology, 2015.

[12] Yue B, Gui M, Guo J. "An effective framework for question answering over freebase via reconstructing natural sequences," Proceedings of the 26th International Conference on World Wide Web Companion. 2017: 865-866, to be published.

[13] Ritze D, Bizer C. "Matching web tables to dbpedia-a feature utility study," context, 2017, 42(41): 19-31, to be published.

[14] Webber J. "A programmatic introduction to neo4j," Proceedings of the 3rd annual conference on Systems, programming, and applications: software for humanity. 2012: 217-218, to be published.

[15] Goldberg Y, Levy O. "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method," arXiv preprint arXiv:1402.3722, 2014.

[16] Pennington J, Socher R, Manning C D. "Glove: Global vectors for word representation," Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543, to be published.

[17] Huang Z, Xu W, Yu K. "Bidirectional LSTM-CRF models for sequence tagging" arXiv preprint arXiv:1508.01991, 2015.

[18] Sutton C, Mccallum A. "An Introduction to Conditional Random Fields," Foundations and Trends in Machine Learning, 2010, 4(4):267-373, to be published.

[19] Wu Z，Jiang D，Wang J，et al. "Knowledge-based BERT: a method to extract molecular features like computational chemists," Briefings in Bioinformatics，2022,23(3):bbac131, to be published.