Intelligent Information extraction algorithm of Agricultural text based on Machine Learning method

Zihao Wang1,*, Zilong Cheng2, Chenzheng Guan3, Haoxiang Han1

1College of Tropical Crops, Hainan University, Haikou, Hainan 570100
2College of Applied Science and Technology, Hainan University, Danzhou, Hainan 571700
3Nanjing Agricultural University, Nanjing, Jiangsu, 210095

*Corresponding author: wangzihao@hainanu.edu.cn

Abstract. The Internet agricultural technology question and answer platform now only relies on manual to provide answer service, the response speed is slow, and the answer quality is difficult to be guaranteed. In order to realize the intelligent question and answer of agricultural technology and construct the knowledge base of agricultural technology, it is necessary to extract the named entity triple of "crop-pest-pesticide" from the existing question and answer data. There are few researches on agricultural Chinese named entity recognition, and the accuracy is low. According to the characteristics of named entities of crops, diseases and insect pests and pesticides, and according to the question and answer data of agricultural technology, a method of identifying named entities of crops, diseases and pests and pesticides based on conditional random field was proposed. The data set is formatted and segmented automatically, and the corpus after word segmentation is automatically tagged according to whether it contains a specific definition word, whether it contains a specific partial radical, whether it is a quantifier, whether it is a specific left and right definition word and part of speech. Using the tagged data to train the CRF model, we can classify the corpus, including judging whether the corpus belongs to crop, pest and pesticide named entities and identifying the position of the corpus in the compound named entity, thus realizing the recognition of the three kinds of named entities and automatically constructing the associated triple. Through the experiment to select the feature combination and adjust the context window size, the recognition accuracy of this method is improved, the model training time is reduced, and the accuracy of crop, pest and pesticide named entity recognition is 97.72%, 87.63% and 98.05%, respectively, which is significantly higher than the existing methods.

Keywords: diseases and insect pests; Pesticide; knowledge Base; named entity recognition; conditional Random Field.

1. Introduction
With the rapid development of "Internet +", the national agricultural e-commerce industry is showing an explosive growth trend, there are more than 30000 agriculture-related e-commerce platforms [1],
agricultural technology consulting service has become the standard configuration service of most agricultural e-commerce platforms. Farmers can describe crop growth and disease through mobile phone APP, and get answers from platform experts in time. In the face of the geometric growth of question data, the existing platform also relies on manual online answers from domain experts, and timeliness and professionalism have become the deficiency of the level of agricultural technology consulting services. How to construct the knowledge base of "crops, diseases and insect pests-pesticides", so as to support the intelligent solution of agricultural technical problems, has become an important problem.

Based on CRF model, this paper studies the method of named entity recognition of crops, diseases and insect pests and pesticides. Using more than 70 thousand question-and-answer data captured by APP, the agricultural housekeeper, by analyzing the characteristics of the names of crops, diseases and insect pests and pesticides, such as part of speech, partial radical, left and right demarcation words, and nearby quantifiers, the CRF model was trained to establish the relationship between the above features and the categories and morphemes of named entities, so as to identify named entities.

2. Data set

This paper analyzes the question and answer data of farmers and experts in the APP products of Shandong Laodao Network Technology Co., Ltd. on the 1st and 15th of each month in 2016, including: 7870 questions on farmers describing crop diseases and 66559 answers from experts on the platform. Compared with the pesticide catalogue and other data sets, agricultural question and answer data have the advantage of timeliness, by mining "crops-diseases and insect pests-pesticides" information and associated address and other information, it can accurately provide farmers with suggestions on the common use of certain crop diseases and insect pests in a certain place in the near future.

The data items of each question include: question ID, question content, province, number of answers, etc., covering vegetables, food crops, fruit trees and other crops.

The data items of each answer data include: reply ID, question ID, reply content, etc. There is a corresponding relationship between the answer and the question, and the answer can be associated with the question through the question ID.

3. Named entity recognition method based on CRF

3.1. CRF model

In this paper, a named entity recognition method based on CRF is proposed, which is used to identify named entities of crops, diseases and insect pests and pesticides from agrotechnical question and answer data, and to establish a ternary knowledge set of "crops, diseases and insect pests-pesticides". CRF model is a probabilistic undirected graph model used to label and partition sequential structural data, which is widely used in machine learning in natural language processing. As shown in figure 1, for a given input text sequence x, it is obtained by word segmentation and feature tagging \((x_1, x_2, \cdots x_p)\) corpus sequence, the model parameters obtained through training are used to predict the conditional probability of the required corpus tagging combination y.

![CRF model structure](image)

\[
X = X_1, \ldots, X_{n-1}, X_n
\]

Figure 1 CRF model structure

Describe the conditional probability of the CRF model \(P(y | x)\) the calculation formula is:
\[ P(y \mid x) = \frac{1}{Z(x)} \exp \left( \sum_{i=1}^{n} \sum_{j} w_{j}f_{j}(y_{i-1}, y_{i}, x, l) \right) \]

Of which

\[ Z(x) = \sum_{y} \exp \left( \sum_{i=1}^{n} \sum_{j} w_{j}f_{j}(y_{i-1}, y_{i}, x, l) \right) \]

The training process is known to be a corpus set

\[ D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \]

with both input and output tagged. Calculate the parameters of the model. The test process is to find the annotation sequence \( y \) with the highest conditional probability given \( P(y \mid x) \) and observation sequence \( x \).

For the corpus observation set \( x \) after word segmentation, six features are selected, such as whether it contains a specific definition word, whether it contains a specific partial radical, whether it is a quantifier, whether it is a specific left and right definition word and part of speech. Because the language presents the characteristics of chain sequence, the tagging of \( x \) takes into account not only the current corpus features, but also the contextual features of the corpus in order to improve the recognition accuracy.

After testing, the CRF model completed by training can be used to accurately identify three kinds of named entities: crops, diseases and insect pests and pesticides in the question and answer text.

3.2. Data format arrangement and word segmentation

The question and answer data is irregular in the data format and cannot be directly used for processing. The data format needs to be sorted out.

Accurately segment the sentence and cut the sentence into corpus, such as "it's not drug damage, it's bacterial leaf spot, simply use Nonglian plus quinolinone twice to reduce humidity." After investigation, it is found that the existing word segmentation tools are more mature, so I will not repeat them here.

3.3. Automatic feature tagging of observed corpus sequences

This paper selects the defining words (Word), part of speech (Part-of-speech,POS), left and right referential words (Leftbound+Rightbound, LB+RB), partial radical (Radical, Rad) and quantifier (Numeral, Num) as features to label each corpus and its context in the observed corpus sequence. These features have a good degree of differentiation, and it is easy to realize automatic labeling through the program.

First of all, the definition term (Word): crops often contains characteristic morphemes such as "melons, beans, vegetables", diseases and insect pests often contain "diseases and insects" morphemes, and pesticides often contain morphemes such as "phenols, lipids, esters, amines" and so on. These characteristic morphemes become an important basis for identifying named entities of crops, diseases and insect pests and pesticides. However, due to the emergence of "what disease" and "what problem", it is limited to look for characteristic words from the position of the second word symbol, and to rule out such invalid words as "illness" and "illness".

Second, the part of speech POS: crops as nouns, and diseases, pests and pesticides in addition, there are compound nouns, including some verbs, adjectives and so on. Therefore, it is necessary to label the part of speech. For example, "no / d is / vshi drug / n harm / ng/ wd is / vshi bacteria / n sex / ng leaf spot / njingle / wd is simple / udel1 / v farmer / ng lotus / ng plus / v quinoline / none / w / d / v two / m times / qv/, wd decrease / v humidity / n" and remove stop words, such as "thank you", "Mo" and so on. For the meaning of part of speech, see ICTCLAS Chinese part of speech tagging set.

Third, named entities such as LB, RB: crops, pests and pesticides often appear together with specific predicates, verbs and adverbs. Some words appear on the left side of the named entity, which is called the left finger boundary word, and the one on the right side is called the right finger boundary word. The
position of the beginning of a sentence is also a commonly used left finger boundary, and punctuation marks can also be used as the right finger boundary.

4. **Test and result analysis**

Choose ICTCLAS as the tool of word segmentation. CRF++0.58 is a conditional random field tool that can be used for word segmentation / continuous data tagging, and it is also necessary to generate corresponding lexical tagging and category tagging for crop, pest and pesticide named entity recognition. Therefore, this paper uses CRF++0.58 toolkit for training and testing.

4.1. **Single feature annotation**

This experiment examines the accuracy of the NER method in this paper when only one of Word, POS, Num, Rad, LB and RB6 features is used for labeling. The dataset uses a full set of 7870 question data and 66559 answer data. Among them, 75% of the random division is used for model training and 25% for test model.

The test results are shown in Table 1. When only a single feature is used, except for the Word feature, the accuracy of the named entity recognition with other features is low, especially when the feature selection is LB or RB, the recognition effect is very poor because the structure of the text content grammar itself is incomplete. Therefore, it is necessary to use a variety of features together to improve the recognition accuracy.

| Feature selection | Types                      | P   | R   | F   |
|-------------------|----------------------------|-----|-----|-----|
| **Word**          | Crops                      | 64.84 | 63.37 | 64.09 |
|                   | Diseases and insect pests  | 58.38 | 95.92 | 72.58 |
|                   | Pesticide                  | 82.69 | 84.84 | 83.75 |
| **POS**           | Crops                      | 10.23 | 33.33 | 15.66 |
|                   | Diseases and insect pests  | 0    | 0    | 0    |
|                   | Pesticide                  | 68.45 | 15.76 | 25.62 |
| **Num**           | Pesticide                  | 2.913 | 24.90 | 5.210 |
| **Rad**           | Crops                      | 19.23 | 31.13 | 17.83 |
|                   | Diseases and insect pests  | 24.52 | 15.17 | 18.74 |
|                   | Pesticide                  | 24.98 | 15.41 | 19.06 |
| **LB**            | Crops                      | 16.67 | 33.33 | 22.23 |
|                   | Diseases and insect pests  | 0    | 0    | 0    |
|                   | Pesticide                  | 38.71 | 20.67 | 26.95 |
| **RB**            | Crops                      | 0    | 0    | 0    |
|                   | Diseases and insect pests  | 9.71 | 0    | 0    |
|                   | Pesticide                  | 64.84 | 25.02 | 14.01 |

4.2. **Feature combination dimensioning when considering context**

Without considering the context, even if the multi-feature combination is selected, the recognition accuracy of diseases and insect pests and pesticides is still low, which is only about 80% and 90% respectively. Therefore, the consideration of contextual joint tagging is added in this experiment. The dataset uses a complete set of questions and answers. Among them, 75% of the random division is used for model training and 25% for test model. Compare the following four context windows: window A, current word only; window B, current word + one word above; window C, current word + one word below; window D, current word + context one word each.

The test results are shown in Table 2. It can be seen that window C, that is, when considering the current word and the following word, the F value of representation recognition accuracy is the highest,
and the F value of named entity recognition of diseases and insect pests is increased to 87.63% and 98.05% respectively.

| Window | Types                  | P     | R     | F      |
|--------|------------------------|-------|-------|--------|
| A      | Diseases and insect pests | 96.33 | 69.50 | 80.74  |
|        | Pesticide              | 96.05 | 88.56 | 92.15  |
| B      | Diseases and insect pests | 89.76 | 71.11 | 79.35  |
|        | Pesticide              | 99.00 | 95.29 | 97.11  |
| C      | Diseases and insect pests | 90.89 | 84.59 | 87.63  |
|        | Pesticide              | 99.01 | 97.10 | 98.05  |
| D      | Diseases and insect pests | 78.62 | 82.55 | 80.54  |
|        | Pesticide              | 98.42 | 96.02 | 97.21  |

5. Conclusion
The main contents are as follows:
1) A method for identifying named entities of crops, diseases and insect pests and pesticides based on conditional random fields is proposed. Methods the defining words, part of speech, left and right referential boundary words, partial radicals, quantifiers as features, as well as the contextual relationship between corpus were considered.
2) The test shows that the F value of the method for the identification accuracy of crops, diseases and insect pests and pesticides is 97.72%, 87.63% and 98.05%. All of them are higher than the recognition methods of the existing literature.
3) Balance operation performance and recognition accuracy, when the machine performance is weak, it is recommended that the size of the context window is 3 and the number of features is 4. When the performance of the machine is good, it is recommended that the size of the context window is 7 and the number of features is 4.

References
[1] Yu Lianjun. Research on the Development Model of Agricultural Electronic Commerce based on Internet + [J]. Agricultural Network Information, 2015 (11): 19-21
[2] Chen Ji. an overview of named entity recognition [J]. Modern computer, 2016 (3): 24-26
[3] Cai Aijie. Research method of named entity extraction based on Web [J]. Journal of Natural Science of Harbin normal University, 2010.
[4] Tanigawa, Zhou Hongyu, Yu Jiangde. Chinese product named entity recognition based on multi-features [J]. Science, Technology and Engineering, 2013 13 (31): 9417-9421
[5] Jiang Renhui, Wang Ting, Tang Jintao. Named entity recognition for Weibo text [J]. Computer and Digital Engineering, 2014
[6] Li Lishuang, he Honglei, Liu Shanshan, et al. Biomedical named entity recognition based on word representation [J]. Minicomputer system, 2016Magi 37 (2): 302-307.
[7] Feng Yuntian, Zhang Hongjun, Hao Wenning. Named entity recognition for military text [J]. Computer Science, 2015