Identification of Tomato Diseases using Skip-gram and LSTM Based on QA(Question-Answer) System

Xiao Hang, Hongju Gao* and Shaopeng Jia

College of Information and Electrical Engineering, China Agriculture University, Beijing 100083, China

Corresponding author: hjgao@cau.edu.cn

Abstract. In the field of agricultural information processing, automatic identification and diagnosis of common diseases of tomatoes play an important role. Deep learning is a hot research topic in the field of pattern recognition and machine learning. It can effectively solve some problems of vegetable pathology, such as disease identification, automatic control and production prediction. In this paper, the tomato pest and disease experiments were carried out with natural language data set of 1.12G pathological and healthy tomatoes crawled from Doctors of Agriculture Website as source data. Five common tomato diseases' symptoms were identified by training Skip-gram algorithm. Finally, symptoms identified above can be classified using LSTM algorithm with classifiers. The research shows that the tomato pest and disease corpus identification models based on LSTM algorithm with classifiers achieve an average accuracy which are over 60%. The simulation results of tomato diseases identification show feasibility and effectiveness of the methods. This article aims at integrating our methodology into working systems that can be used in the identification fields.

1. Introduction

Approximately 20%-30% of tomato production is lost by attack of pests, diseases and nutritional disorders of plants, even up to 50%-70% when prevalent years occur. A rapid and efficient diagnosis of these situations is essential to avoid greater losses. It is estimated that 10%-15% of the damage is caused by error diagnosis and delay diagnosis. When uncertainty is in a high degree, diseases often occur. Among these diseases, botrytis cinerea, late blight, leaf mold, early blight and viral diseases are the most important ones. In the case of pests, the urgent issues are due to the attack of fly, moth and caterpillars. On the other hand, the disturbances caused by the diseases are visible in early time during vegetative growth phase, which is more important and prevalent.

A correct diagnosis is essential in order to define strategies of management and control, and consequently for the rational use of fertilizers and pesticides. One main obstacle towards a quick and accurate diagnosis system is the need for experts, making it costly to cover large domains and give a feasible solution in time. Moreover, experts often specialize in specific issues, increasing the rate of misdiagnosis. Some attempts have been tried to reduce the dependency on experts. Expert Systems, often built on top of Case-Based Reasoning algorithms, are applied to some culture. These systems still require considerable amounts of training and are often not accurate, mainly due to the typically very large number of questions required to be answered by experts and the sensitivity to the wrong answers.

The concept of using machine learning to detect symptoms of plants has been shown to be a promising alternative in recent years, where several studies using different approaches have been
carried out to identify or classify symptoms of plants. Dai at al. reviewed on how the Point-of-care (Poc)-associated nanotechnology currently applied to the identification of nucleic acids, proteins and antibodies [1]. Singh G et al. proposed an innovative framework that uses principal component analysis (PCA) for feature extraction, Fisher discriminant ratio (FDR) for feature selection and support vector machines (SVM) for classification of Healthy controls, Parkinson Disease and SWEDD subjects to diagnose neurodegenerative diseases (NDDs)[2]. They have extended their framework to handle the challenge of multi-class disease diagnosis, wherein, accuracy up to 100% has been achieved. Nurlaeli et al. focused on chaining method in the diagnosis diseases and pests of corn crop to help farmers/agricultural facilitators in getting knowledge about disease and pest corn crop [3]. The percentage accuracy of results of diagnosis using data tested is 88%.

Early diagnosis and treatment are the most effective means of ensuring biological health. Now, with the advent of more and more expert systems, on one hand, it promotes the collection of materials and people can get useful information from it; on the other hand, the information is more complicated and difficult for users to identify, such as inappropriate words and incomplete words.

Byrd et al. developed a natural language processing procedure to identify Framingham HF signs and symptoms among primary care patients, using electronic health record (EHR) clinical notes, as a prelude to pattern analysis and clinical decision support for early detection of HF [4]. The criteria mention extractions achieve a precision of 0.925, a recall of 0.896, and a F-score of 0.910. Encounter labeling achieves a F-score of 0.932. The system accurately identifies and labels affirmations and denials of Framingham diagnostic criteria in primary care clinical notes and may help in the attempt to improve the early detection of HF.

Information gathering from patient by clinicians during diagnostic procedures may sometimes require some skills to adequately collect required information that will be sufficient for the procedure. Oyelade et al. proposes a formalized input generating model that addresses this shortcoming through the creation of an inference process, breast cancer lexicon, rule set and natural language processing (NLP) [5]. They developed an input generation algorithm which uses the python natural language processing capability in first filtering and generation the first pre-input collection. This proposed model was tested on a breast cancer based DDSS earlier designed by the authors, and result shows that the inference supporting this model generates additional input of about 64%.

LSTM has been shown to be a good alternatives in the most difficult problems of natural language processing, such as sentiment classification [6], time series classification [7], text recognition [8] and sentence classification [9].

The combination of word segmentation, Skip-gram algorithm and LSTM provides a more accurate basis and result to characterize diseases of tomatoes. In this study, in order to simplify the sentence, word segmentation and vector representation of words are used. Then, Skip-gram algorithm is used to find the most similar words, and the relationship between related words which provides a method used to analyze and identify the types of pests and diseases. Finally, LSTM is applied to reproduce sentences to attain the exact symptom of the plants and several classifiers are then trained to get the real diseases. The main objectives of this paper are: (1) to reproduce the sentences people uploaded to make them more complete and comprehensible; (2) to identify specific types of diseases in tomatoes; and (3) to evaluate accuracy of the proposed method with different classifiers.

2. Materials and methods

2.1. Data set

The data set were built by crawling natural languages from online website(www.nongyisheng.com) and then diseases’ materials about tomatoes were selected from data above. Each language material was diagnosed referencing experts’ diagnoses and users’ suggestions to identify the underlying disorder which was then used to label the language material. Data were collected from December of 2017 to May of 2018.

Five disorders were selected among those collected which are shown in Table 1, as they are the
most prevalent in the website above. Selected symptoms represent four damages caused by high humidity, which are botrytis cinerea, early blight, late blight and leaf mould diseases, respectively. Viral diseases of tomato happen mostly when temperature is high. On the contrary, botrytis cinerea is prevalent while temperature is low. Table 1 also identifies the number of samples collected for each disease.

The data used in this study consist of question narratives and answer narratives whose lengths are shown in Table 2 and Table 3, respectively. On total, answer narratives are longer than those of question narratives. As to average length, the character length of answers is about three times longer than those of questions. And the maximum length of questions is 262, half of the length of answers roughly.

The identification of disorders was conducted by a group of professional experts specialized in these symptoms and with ample experience in vegetable nutrition and vegetable pathology, as well as experienced farmers. Also, to further support these professionals, published tomato books were used, providing technical support on symptoms.

2.2 Identification system architecture

Figure 1 describes all processing procedure of algorithms to connect questions and advice with natural language processing algorithms which aim at identifying pest and diseases of tomatoes as well as investigating the relationships between symptoms and diseases. This process consists of database creating and diseases identification which includes word segmentation, investigation of relationships between symptoms and diseases based on word vectors and LSTM with classifiers.

![Figure 1. Processing procedure for pest and diseases identification.](image)

2.3 Database Creating

The process of database creating consists of two steps: firstly, crawling pest and diseases information of tomatoes; secondly, labeling each natural language material of tomato diseases. Since natural language materials uploaded are informal, after crawling all data in the certain period from websites, natural languages about tomatoes were selected specifically and then label each natural language material synthesizing the analysis of experts, farmers and books. Finally, stored to database in order. A few natural language materials which presented defects or were out of investigation were discarded.

Table 1 illustrates five kinds of diseases selected in this paper, with conditions where they happen mostly and number of language materials each disease set contains. Also samples of language materials crawled in this study are shown in Table 4, which include questions, answers, time and labeling results.

| Numbers | Diseases            | Conditions                      | Collected number |
|---------|---------------------|---------------------------------|------------------|
| 1       | Botrytis cinerea    | Low temperature and high humidity | 6890             |
| 2       | Late blight         | High humidity                   | 6840             |
2.4 Disease Identification

2.4.1 Word Embedding and Symptom Identification.
After word segmentation, word representation is essential to performance of symptom identification algorithms. Word Embedding which is proposed to simplify the NNLM (Neural Network Language Model) was used to train word vectors, and then vector representation of the whole document was attained by word weighting, key tag weighting and tf-idf (term frequency-inverse document frequency) weighting. Word2vec is a popular and effective method among Word Embedding. Skip-gram is one of word2vec models, which is applied in this paper.

In this paper, Skip-Gram model based on Hierarchical Softmax was used to train the Word Embedding model. The embedding size (dimension of the embedding vector) is set to 150. The batch_size (number of pieces per scan) varies from 36 to 176, step by 4. Embedding_size converts the word into a dimension of a dense vector, which is usually a range (50-1000). Experiments have shown that a better dimension of the word vector in this paper is 150, which can be seen clearly in Figure 2; skip_window (how many words to consider left and right) is the farthest distance a word can reach, which is set to 1; num_skips is the number of samples to be extracted for each target word, set to 4. Then, this article generates valid_examples, where some of the words with the highest frequency are randomly selected to be viewed in vector space, and the most recent words are more relevant; Valid_size indicates the number of validation words to be extracted and is set to 16, and valid_window is set to 50, indicating that validation words can only be extracted from 50 words with the highest frequency. This article defines SGD (stochastic gradient descent) as the optimizer.

The similarity between words, similar words to a certain word and the inconsistent word among word set can also be found out, which could be helpful in assisting understanding the need among people, targeting argent problems, and investigating connection between features and diseases.

2.4.2 LSTM with classifiers.
Furthermore, since the words uploaded by people with different education background, occupations and habits, missing words, adding words and colloquial words appear which are obstacle to intelligent identification. LSTM is applied to reproduce complete and comprehensive language materials by training and testing the raw materials. LSTM is a variant of RNN, optimizing vanishing and exploding gradient problem. As the language material processed is usually small and casual, step (how many words to consider left and right) varies from 4 to 44 to find a suitable one. And the values of diversity (sample the results again to get different answers) vary in 0.2, 0.5, 1.0 and 1.2 to make the results more valid.

Finally, several methods are used to classify the processed materials. As for language data set, 34440 language materials were divided into two subsets. The test subset were created by randomly choosing 3/11 of each class (939 examples, 3/11). The remaining examples were in training set.

The tested learning algorithms were trained over the training set using different parameters and configurations. The best performing parameters of each algorithm were then applied to training examples. This procedure was adopted in order to avoid over fitting. The test set was also labeled by referencing the suggestion of experts which are helpful for classifications, as well as classification books. SVM (e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien), LOGITBOOST, BAGGING (ipred: Improved Predictors), and MAXENTROPY (maxent: Low-memory Multinomial Logistic Regression with Support for Text Classification) were used for classification. The evaluation of the results was done by calculating

|   | Leaf mould Diseases | High temperature and high humidity |   |
|---|------------------|-----------------------------------|---|
| 3 |                  |                                   | 6880 |
| 4 | Early blight     | High humidity                     | 6960 |
| 5 | Viral diseases   | High temperature                  | 6870 |
PRECISION (a ratio which are classified correctly), RECALL (a ratio which is classified correctly retrieved accounting for the proportion that should be retrieved.) and F-SCORE (a criteria synthesized values of precision and recall, which can be calculated by \((2 \times \text{precision} \times \text{recall}) / (\text{precision} + \text{recall})\)).

And, to improve the performance of classifier, four classifier algorithms above are combined. When n-ENSEMBLE COVERAGE (n classifiers are contained) varies from 1 to 4, the values of n-ENSEMBLE (a percentage which is the number of materials attaining recall above to all) and RECALL illustrated in Table 5.

Table 2. The character length for the questions

| Minimum | Average | Maximum |
|---------|---------|---------|
| 0       | 21.3168995 | 262     |

Table 3. The character length for the answers

| Minimum | Average | Maximum |
|---------|---------|---------|
| 1       | 70.54122873 | 420     |

Table 4. Samples of tomato Natural Languages

| Number | Questions                                                                 | Answers                                                                 | Time          | Label |
|--------|---------------------------------------------------------------------------|-------------------------------------------------------------------------|---------------|-------|
| 1      | What's wrong with the small potatoes with leaves aging, persimmon dropping and white spots on poles in greenhouse? | It may be Botrytis cinerea and leaf blight which are caused by recently watered. The humidity in the shed is too high. | 3/13/2018 20:51:16 | 1     |
| 2      | Greenhouse tomatoes. Teachers, please see what's wrong with it which also drops tomatoes? What medicine should use? | The picture of tomatoes in greenhouse is blurred. From the symptoms of the back leaves, it is mostly caused by gray mold infection. Pay attention to ventilation, ventilation and dehumidification. | 3/14/2018 19:34:38 | 1     |
| 3      | What kind of disease is it? It develops quickly.                          | It is tomato late blight. It is recommended to spray alternatively with two different fungicides, once every 5 days. | 3/16/2018 16:55:19 | 2     |

Table 5. N-ENSEMBLE and RECALL when N-ENSEMBLE COVERAGE varies from 1 to 4

| N-ENSEMBLE COVERAGE | N-ENSEMBLE | RECALL |
|---------------------|------------|--------|
| N>=1                | 1.00       | 0.44   |
| N>=2                | 1.00       | 0.44   |
| N>=3                | 0.77       | 0.57   |
| N>=4                | 0.77       | 0.57   |

Table 6. Precision of some classifiers with different number of materials

| Model/number | SVM_CLASSIFY | BOOSTING_CLASSIFY | BAGGING_CLASSIFY | MAXENT_CLASSIFY |
|--------------|--------------|-------------------|------------------|-----------------|
| 500          | 0.125        | 1.265654e-14      | 0.48             | 0.125           |
| 2000         | 0.004        | 2.220446e-16      | 0.16             | 0.013           |
| 8000         | 0.004        | 0.02              | 0.24             | 0.042           |
| 12000        | 0.004        | 0.357             | 0.2              | 0.047           |
| 35000        | 0.004        | 0.996             | 0.119            | 0.950           |
3. RESULTS AND ANALYSIS

3.1. The results of sentence reproduced using LSTM

Figure 2 shows the variety of loss along training progress. On total, as steps increase, loss decreases except some points in steps of 8000 and 10000, which are over fitting. Figure 3 presents variety of LSTM loss when different steps are set for training and testing natural languages which vary from 4 to 44, step by 4. The graph shows a logarithmic decrease on loss when training process goes on the whole. In general, loss goes down with the decrease of steps, except the condition where the step is 12. As all known, the uploaded language materials are usually colloquial and casual, when the number of words is too small, it may be difficult to understand their meaning, just as the variety of first two steps. As the number of steps increases, loss decreases. But it is not true that the more language material, the better people understand. So, step 12 is boundary where the number of words is between too lager to implement and a little small to understand. So, it is important to investigate a suitable number of step to understand the meanings better: not only too small to understand meaning, but also too huge for performance of test.

3.2. The results of LSTM with classifiers

From Table 5, it can be seen the variety trends of n-ENSEMBLE is opposite with that of RECALL. And when n-ENSEMBLE is 1 or 2, the value of n-ENSEMBLE reached the highest, while the value of RECALL is lowest. According to Table 5, performance of combined classifier acts better.

Table 6 shows the precision of some models when the number of language material varies from 500 to 35000. The overall trends of precision as number increases are going up, except BAGGING_CLASSIFY. As the number of language material increases, the precision of Boosting_CLASSIFY and MAXENT_CLASSIFY are both arise, which reach 0.996 and 0.950 in the end, respectively. And the precision of some models is static in the end process, such as SVM_CLASSIFY, which implies that they reach the saturation point at those moments. Finally, the precision of BAGGING_CLASSIFY decreases in the end. It means that they reached the best performance in the process of variety to the biggest number.
4. Conclusion
In this paper, we proposed a LSTM network model with some classifies based on word vectors to assist on the task of identification and classification of tomato diseases from natural languages. Based on word vectors, some relevance discoveries are also investigated which are helpful in symptoms identification, environment acknowledgement and diseases identification. We used a raw data set containing language materials of 5 common diseases which are colloquial and casual, all confirmed by experts.

We conclude that the LSTM network model with some classifies provides a useful option for this task, with more robust classifications than other identification algorithms. In this sense, an automated system based on the trained model could contribute towards diagnosis reliability and cost reduction. This also allows for the general approach to be used in different targets for different education backgrounds, occupations and habits. This is important to allow for the robustness and automatic improvement in the model when language materials are made casually and informally.

Compared to previous works, our approach does not require the language produced in laboratory environments or with restrict to users. The LSTM model is able to learn relevant features and synthesize context around, to which we attribute the improved performance.

References
[1] Nhan T D T, Wang H, Sigit S, et al. (2016) Advances in nanomaterials and their applications in point of care (POC) devices for the diagnosis of infectious diseases. Biotechnology Advances., 34(8): 1275-1288.
[2] Singh G, Vadera M, Samavedham L, et al. (2016) Machine Learning-Based Framework for Multi-Class Diagnosis of Neurodegenerative Diseases: A Study on Parkinson’s Disease. Ifac Papersonline., 49(7): 990-995.
[3] Nurlaeli S. (2017) Forward chaining method on diagnosis of diseases and pests corn crop. In: International Conference on Education, Concept, and Application of Green Technology. Semarang. pp.020038-1 - 020038-9.
[4] Byrd R J, Steinhubl S R, Jimeng S, et al. (2014) Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. International Journal of Medical Informatics., 83(12): 983-92.
[5] Oyelade O N, Obiniyi A A, Junaidu S B, et al. (2017) Patient Symptoms Elicitation Process for Breast Cancer Medical Expert Systems: A Semantic Web and Natural Language Parsing Approach. Future Computing & Informatics Journal., 3(1): 72-81.
[6] Guozheng Rao, Weihang Huang, Zhiyong Feng, et al. (2018) LSTM with sentence representations for document-level sentiment classification. Neurocomputing., 308:49-57.
[7] Karim F, Majumdar S, Darabi H, et al. (2018) LSTM Fully Convolutional Networks for Time Series Classification. IEEE Access., 6(99):1662-1669.
[8] Breuel T M. (2017) High Performance Text Recognition Using a Hybrid Convolutional-LSTM Implementation. In: Iapr International Conference on Document Analysis and Recognition. Kyoto. pp.11-16.
[9] Ding Z, Xia R, Yu J, et al. (2018) Densely Connected Bidirectional LSTM with Applications to Sentence Classification. arXiv preprint arXiv:1802.00889.