Question Similarity Measurement of Chinese Crop Diseases and Insect Pests Based on Mixed Information Extraction

Han Zhou[1], Xuchao Guo[1], Chengqi Liu[1], Zhan Tang[1], Shuhan Lu[2], Lin Li[1,*]
[1] College of Information and Electrical Engineering, China Agricultural University
Beijing, 100083, China.
[Email : szhzh@cau.edu.cn, gxc@cau.edu.cn, cieelcq@cau.edu.cn, tz_blues@163.comm, lilincau@126.com]
[2] University of Michigan, Ann Arbor, USA
[Email : shuhanlu@umich.edu]
[*]Corresponding author: Lin Li

Received April 5, 2021; revised August 23, 2021; accepted Septebmer 14, 2021; published November 30, 2021

Abstract

The Question Similarity Measurement of Chinese Crop Diseases and Insect Pests (QSM-CCD&IP) aims to judge the user's tendency to ask questions regarding input problems. The measurement is the basis of the Agricultural Knowledge Question and Answering (Q & A) system, information retrieval, and other tasks. However, the corpus and measurement methods available in this field have some deficiencies. In addition, error propagation may occur when the word boundary features and local context information are ignored when the general method embeds sentences. Hence, these factors make the task challenging. To solve the above problems and tackle the Question Similarity Measurement task in this work, a corpus on Chinese crop diseases and insect pests (CCDIP), which contains 13 categories, was established. Then, taking the CCDIP as the research object, this study proposes a Chinese agricultural text similarity matching model, namely, the AgrCQS. This model is based on mixed information extraction. Specifically, the hybrid embedding layer can enrich character information and improve the recognition ability of the model on the word boundary. The multi-scale local information can be extracted by multi-core convolutional neural network based on multi-weight (MM-CNN). The self-attention mechanism can enhance the fusion ability of the model on global information. In this research, the performance of the AgrCQS on the CCDIP is verified, and three benchmark datasets, namely, AFQMC, LCQMC, and BQ, are used. The accuracy rates are 93.92%, 74.42%, 86.35%, and 83.05%, respectively, which are higher than that of baseline systems without using any external knowledge. Additionally, the proposed method module can be extracted separately and applied to other models, thus providing reference for related research.

Keywords: Text semantic similarity, Short text-similarity, Agricultural natural language processing, Chinese word segmentation

1. Introduction

Crop diseases and insect pests are some of the main agricultural disasters in China. These occurrences have numerous types, great influence, and frequent outbreaks. For example, in 2019, a rice blast broke out in Jianhu County, Jiangsu Province, which covered an area of 21600km^2 and accounted for 51.7% of the total area of rice [1]. In recent years, China's agricultural development has gradually shifted toward organic agriculture. This change requires early detection, correct diagnosis, less application of pesticides, and appropriate treatment to reduce pesticide pollution. Therefore, the development of an accurate and efficient crop knowledge question and answer (Q & A) system for disease and insect pest control, the correct diagnosis of crop diseases and insect pests, and the formulation of effective control strategies are particularly critical. Hence, the Question Similarity Measurement of Chinese Crop Diseases and Insect Pests (QSM-CCD&IP) has been designed to measure the user's intention to ask questions, such as "如何控制病害" (How can diseases be controlled?) and "苹果开裂的原因是什么" (What causes apple cracking?). This measurement is a basic task to construct the Knowledge Q & A system for crop diseases and insect pests [2].

Text similarity measurement has been widely used in electronic commerce [3], social dimension [4], and biology [5]. However, the QSM-CCD&IP still has some challenges. Firstly, the QSM-CCD&IP domain text has numerous specific words, such as "苯醚甲环唑" (a drug name) and "二九南一号A" (a crop name). Such specific terminology makes the migration of existing models a challenge. Secondly, some entity nouns have lexical nesting. In addition, the numbers, letters, and special characters are mixed, and the length of nouns is long, such as "乙螨·螺螨酯40%悬浮剂" (a drug name). These characteristics are a challenge to the model's ability to recognize lexical boundaries. Third, the lack of corpus in the field of QSM-CCD&IP increases the difficulty of related research. At present, only Zhao [6], Zhang [7] and Jin [8] have studied question classification in the field of Chinese agriculture. However, the method is based on the idea of text classification. The data are classified according to the existing categories, and the richness of question intention expression is ignored; the question does not belong to the existing categories, but the intention of the existing categories is present.

In the early days, researchers used methods such as edit distance [9-10], TF-IDF [11], BM25 [12], and vector space model (VSM) [13] to match texts. However, these approaches mainly solve the problem of literal similarity. Given the richness of Chinese meanings, the semantic similarity between two sentences is difficult to determine directly through keyword matching or shallow model based on machine learning. Then, researchers put forward a method to measure text similarity by combining grammatical information [14-15]. However, this method is suitable for texts with a more standardized syntactic structure, and it is not effective for sentences with missing grammatical elements. For example, the sentence "番茄黄" (tomato is yellow) has two intentions: "symptom prevention" and "symptom cause," which is similar to the expression "果树落叶怎么办" (what to do with fruit trees and fallen leaves). However, the former lacks grammatical elements, which easily leads to the wrong judgment of the model. Notably, text classification is a way to measure sentence similarity. Kim et al. [16] applied convolutional neural networks (CNN) to the text classification task (TextCNN) and achieved good results in sentiment analysis and problem classification datasets. In addition, Johnson et al. [17] proposed deep pyramid convolutional neural networks (TextDPCNN) to improve the TextCNN, which cannot capture long-distance dependence. Their method achieved better results in emotion analysis and other tasks; however, the model still has insufficient ability to perform hidden feature mining. Meanwhile, Google [18] put

forward the bidirectional encoder representations for transformers (BERT) in 2018, which is based on the multi-layer transformer structure and can better capture the feature representation of characters. The implicit knowledge that can be learned from a large unlabeled corpus improves the performance of the text classification model. Subsequently, given the excellent performance of the pre-training model in NLP tasks, a series of improved models based on BERT has been proposed, such as XLNet [19], RoBERTa [20], and ALBERT [21]. Although these models can capture sentence context information through the deep neural network, they have a large number of parameters. For example, BERT has 340M parameters. Hence, the size of computing resources limits the application of this kind of model.

The deep semantic matching model can achieve better performance by projecting different texts into a common low-dimensional semantic space for similarity measurement. Yin et al. [22] proposed a Siamese network attention-based convolutional neural network (ABCNN) based on shared weight, which encodes the input sentences by shared weight and then measures the similarity. Although this method can reduce the learning parameters, it makes less use of the information interaction between two sentences. Furthermore, Chen et al. [23] proposed an enhanced sequential inference model (ESIM), a sequential reasoning model based on chain long short-term memory (LSTM). This model optimizes the sequential reasoning model by adding local reasoning and reasoning combinations, and achieves the best results on the SNLI with an accuracy of 88.6%. Meanwhile, Wang et al. [24] proposed bilateral multi-perspective matching (BIMPM), which can capture the interaction characteristics between two sentences by matching the independent context information unit between two sentences. The experimental results on Quora reveal that the interaction between sentences can effectively improve the model's ability to measure text similarity. However, BIMPM has several parameters and a slow training speed. Mirakyan et al. [37] proposed a densely interactive inference network (DIIN), which can obtain a high-level understanding of the question pairs through extracting the semantic features using dense interaction tensors (attention) network, and achieves a better result on the Quora with an accuracy of 89.06%. Yang et al. [25] proposed the RE2 model based on enhanced residual information and achieved good application results on Scietail and Quora Question Pairs2. Although the existing Siamese network model can distinguish the similarity of the intention between corpora through information interaction, the ability to extract the existing information of short text is still insufficient, thus leading to errors in the coding layer being transferred to the subsequent network layer and resulting in errors in the model processing results.

The problems of the QSM-CCD&IP are as follows: (1) lack of data set, (2) insufficient ability to extract local information, and (3) difficulty recognizing word boundaries in sentences. Therefore, we have collected and constructed a Chinese crop pest similarity measure corpus (CCDIP). This corpus involves 13 types of questions and has 10559 labeled sentences. At the same time, this work proposes a Siamese network model named the Chinese question similarity of agricultural diseases and insect pests (AgrCQS) which is based on mixed embedded information. The model takes the Siamese network as its basic framework. Then, it uses multi-source word segmentation tools proposed in this study to divide the word boundary and reset the character vectors in the vocabulary by weighted method. The model also combines character-level and vocabulary-level information. Furthermore, it extracts the multi-scale local semantic information of sentences using the MM-CNN proposed in this study. Then, it strengthens the extraction ability of text information and captures the semantic dependence of sentences by the self-attention mechanism. At the same time, the network layer used in this study can be migrated to the existing model. This element provides a reference for building a correct crop diseases and insect pests Q & A system and designing new models in the future.

2. Materials and Methods

2.1 Construction of CCDIP

To ensure the quality of the corpus, we obtain the original information from large Q & A websites, such as hot agricultural investment networks[1] and plant protection technology networks[2]. From these websites, we acquire 25413 pieces of sentences. Then, the raw data are preprocessed, such as encoding conversion and duplicate text deletion. Finally, 19327 question sentences are obtained. To reduce the labor cost and time cost in the process of data annotation, this study uses the combination of keyword clustering and manual verification to annotate the obtained text. The annotation process of the corpus is shown in **Fig. 1**. Data set construction is divided into three stages: keyword clustering, manual verification, and corpus generation.

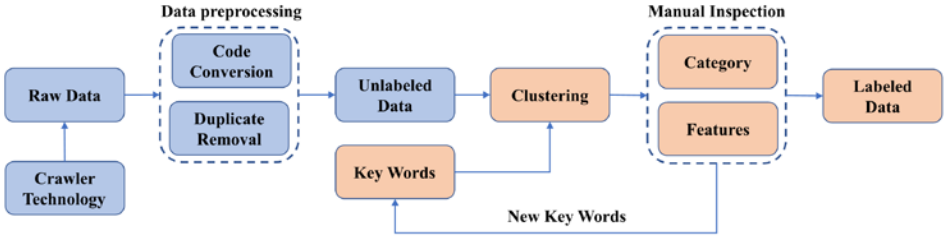


Fig. 1. Labeling process

(1) Keyword clustering

This study constructs a keyword dictionary, which mainly contains words related to the Q & A field of crop diseases and pests, such as diseases, pests, and question intention keywords. Building a keyword dictionary helps to maintain the consistency of question classification. At the same time, expanding the new feature words confirmed in the manual verification process to the dictionary can improve the accuracy of clustering. In the process of clustering, coarse-grained grouping is initially carried out according to the keywords, such as “disease,” “pest,” and “control.” Then, the term frequency-inverse document frequency (TF-IDF) algorithm is used to cluster the answers corresponding to the question in the current group.

(2) Manual verification

After clustering, question clustering error may still occur in the text. Therefore, this study uses manual verification to check the annotation results after keyword clustering and adds the new keywords to the keyword dictionary. After labeling part of the data, the remaining text is re-clustered and verified manually. Finally, 28 different problem classifications are obtained.

(3) Corpus generation

On the basis of the basic semantic basis of the “diagnosis and control of crop diseases and insect pests,” this research removes or merges the question categories with a small number of sentences. Finally, the corpus is divided into 13 categories. Abbreviations and examples of question categories are shown in **Table 1**. In this work, the original corpus is divided by keyword clustering, and the clustered text is checked manually. Finally, 10559 labeled question sentences are obtained.

[1] http://ask.3456.tv/all/1
[2] http://www.zgzbao.com/news.asp

Table 1. Abbreviations and examples of question categories

Id	Abbreviation	Type	Example
1	DIS-IN	Disease information	苍术根腐病有什么症状 What symptoms does Atractylodes root rot have?
2	CAU-DS	Causes of disease or symptom	辣椒树叶子为什么卷曲 Why do the leaves of pepper curl?
3	PRE-DS	How to prevent the disease or symptom	茄子青枯病该怎样防治 How can eggplant bacterial wilt be controlled?
4	ROU-DIS	Route of disease transmission	百香果叶斑病的传播途径 What is the transmission route of a passion fruit leaf spot?
5	PATH-CD	Pathogens causing diseases	山楂根朽病的病原 What is the pathogen of Hawthorn root rot disease?
6	PEST-IN	Insect pests information	银杏大蚕蛾有什么危害 What harm does the ginkgo big silkworm moth bring?
7	OC-PEST	Occurrence factors of insect pests	蔬菜根结线虫的发生原因 What is the cause of vegetable root-knot nematode?
8	CON-PEST	How to control pests	银杏大蚕蛾有哪些防治方法 What control method does the ginkgo big silkworm moth have?
9	PEST-MA	Integrated pest management	苹果开花前后病虫害如何防治 How can apple diseases and insect pests be controlled before and after flowering?
10	PHAR-IN	Pharmaceutical information	哒螨灵怎么使用 How is Pyridaben used?
11	CORP-DP	What diseases and insect pests do crops have	种植向日葵要防治哪些病害 What diseases should be controlled when planting sunflowers?
12	IDE-PRE	Identification information of disease or symptom and how to prevent it	黄瓜绿斑驳花叶病的症状与防治方法 What is the symptom of cucumber green mottle mosaic and how can it be to controlled?
13	CAU-PRE	The cause of the disease or symptom and how to prevent it	大樱桃落花落果的原因是什么？有什么应对措施 What causes falling flowers and fruits in big cherries? What are the countermeasures?

On the basis of the principle that the question sentences appear only once, the CCDIP containing 5100 sentence pairs are generated by randomly pairing different types of statements. The CCDIP is divided into the training set and the test set according to the ratio of 4:1. The question statistics of each category in the training set and the test set are shown in **Fig. 2**. We demonstrate the rationality and validity of the CCDIP in the discussion section.

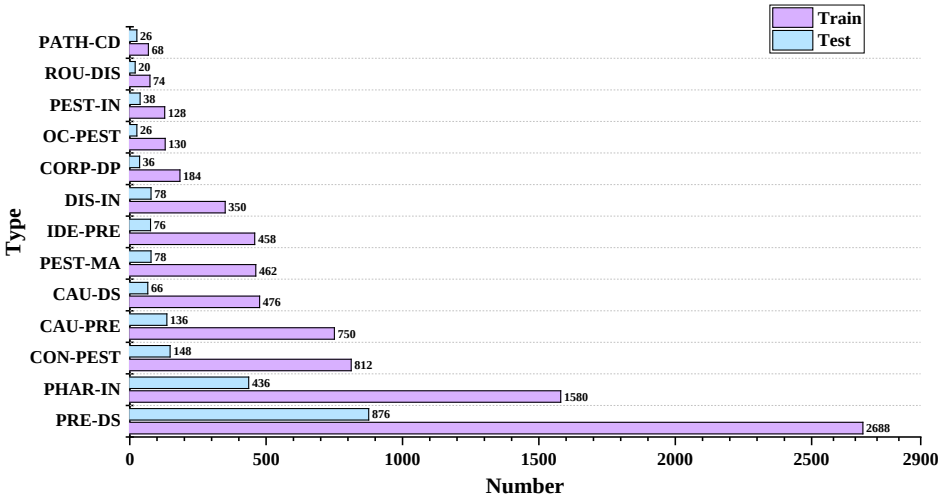


Fig. 2. Statistics of questions in CCDIP

2.2 Analysis of Corpus Features

The CCDIP contains numerous proper terms, such as diseases, pests, and pesticides, which are different from the general domain corpus in categories, specialty, and diversity of professional terms, such as LCQMC [26], BQ corpus [27], and AFQMC [28]. These terms are described as follows:

(1) LCQMC is a question semantic matching dataset built by the Harbin Institute of Technology in COLING 2018. The goal of the task is to search questions that have similar intent as the input question from an existing database. The corpus consists of over 260000 question pairs.

(2) BQ corpus is a large-scale domain-specific Chinese corpus for sentence semantic equivalence identification. It contains 120,000 question pairs from online bank custom service logs.

(3) AFQMC is a financial semantic similarity dataset constructed by Ant Group. The dataset includes three parts: training set (34334), verification set (4316), and test set (3861). This research has only obtained the training set and the verification set. Therefore, in the subsequent research, the data are re-divided for subsequent experiments.

The detailed statistical information of the three datasets is shown in **Table 2**.

Table 2. Comparison information table of each dataset

Datasets	Examples	Size	Max length	Avg length	Pos: Neg
BQ Corpus	今天放款今天就得还款？ Today's loan, today's repayment?	120k	153	11.9	1:1
LCQMC	手机学日语的软件 What is the software for learning Japanese by mobile phone?	260.07k	131	10.73	1.35:1
AFQMC	用花呗怎么买不了车票 Why can't I buy a ticket with Huabei?	38.65k	112	26.62	0.45:1
CCDIP	菠菜叶子发黄怎样补救 How to remedy the yellowing of spinach leaves?	5.1k	34	26.01	1:1

(1) The domain-related words in the CCDIP lack clear boundary features, such as “稻粒瘟” (rice blast) and “枯萎” (wilt). Identifying the boundaries of these words challenges the ability of the model. The conventional word segmentation tools have a poor recognition effect on these words and can easily lead to the wrong transmission. Therefore, using word segmentation tools to divide word boundaries directly is not suitable for the QSM-CCD&IP.

(2) The CCDIP has abundant domain terms. However, in the financial field, domain-related words are more repetitive. For example, “借呗,” “花呗” and “蚂蚁” often appear in the BQ corpus and the AFQMC. Given that these words are highly repetitive, they can be identified by dictionaries. The CIDQS has numerous domain-related words. For instance, “稻瘟病” (disease name) has four aliases: “稻热病,” “火烧瘟,” “吊颈瘟,” and “叩头瘟, ” which makes the dictionary-based method insufficient for identifying all entity nouns. Moreover, the common word vector embedding table has insufficient coverage of domain words. Hence, the model poses a challenge in the representation of word embedding.

(3) The average length of the CCDIP is 26.01. However, the maximum sentence length is smaller. This observation shows that compared with other corpora, the length of sentences in the CCDIP is approximately 26.01. Therefore, the length of sentences in the CCDIP is generally longer; consequently, the model must be able to capture the semantic dependency information of sentences.

(4) The CCDIP has a large number of questions with multiple intentions, such as “造成棉花烂铃的原因以及解救措施有哪些?” (What causes cotton boll rot and what are the rescue measures needed to combat it?). This question expresses the hope to understand the causes of cotton boll rot and how to deal with its symptoms. The model must be able to capture local context information and identify multiple intentions expressed by such questions.

3. Research Methods

Aiming at the problems of the QSM-CCD&IP and the data characteristics of the CCDIP, this study proposes a hybrid embedding layer to encode sentence text to solve the problems of difficult word boundary recognition and insufficient lexical information embedded in characters. The model uses an MM-CNN layer to extract the local context information of different sizes to remedy the weak ability to capture context information. The self-attention layer is used to enhance the ability of the model to extract core sentence meaning and address the insufficient long-distance dependence of BiLSTM [29]. On the basis of the above methods, this study constructs a Chinese agricultural text similarity matching model AgrCQS, which takes ESIM as the basic architecture. The structure is shown in **Fig. 3**. This chapter will focus on the implementation details of the hybrid embedding layer and MM-CNN layer. Soft-attention and BiLSTM are implemented with reference to the literature [30-31].

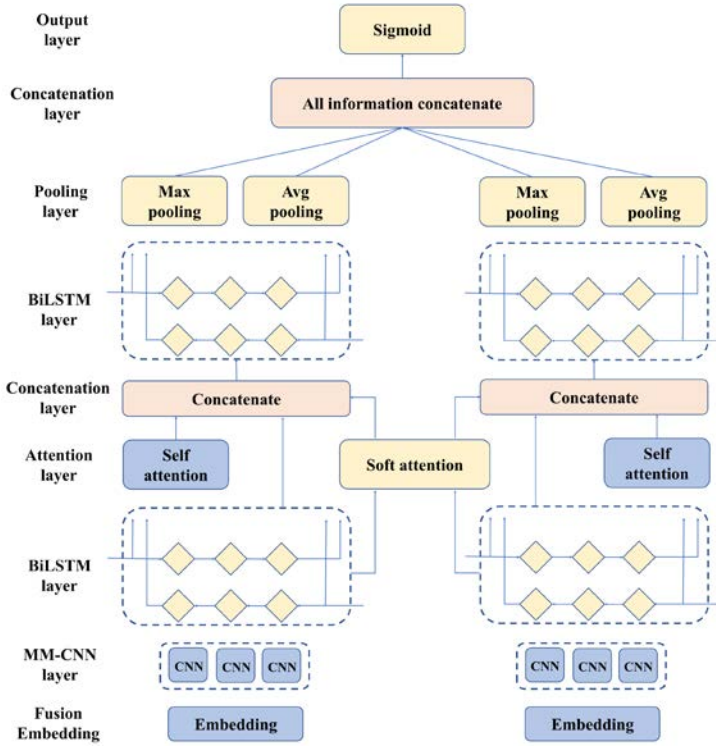


Fig. 3. Structure diagram of AgrCQS model

3.1 Hybrid Embedded Layer

The conventional sequence annotation model and Chinese word segmentation tools are prone to error propagation when recognizing domain word boundaries. Moreover, the traditional character embedding information fails to consider the influence of character position. Thus, this research proposes a hybrid embedding layer to solve the problem. The structure is shown in **Fig. 4**. Considering that the corpus used in this study is small and fails to reach the amount needed for model fine-tuning, we use the Word2vec [32] pre-training vector as the input of the embedded layer of the model.

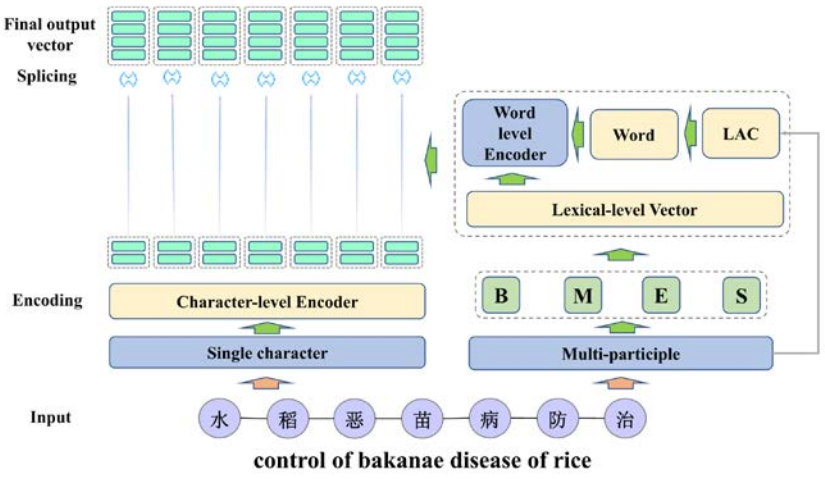


Fig. 4. Structure of hybrid embedding layer

In the previous part, given the input sentence $\{S_n\}_{n=1}^{N}$, the length of the sentence is n. Then, the model uses the word segmentation tool to process the input sentence and obtain some words, that is, $\{W_m\}_{m=1}^{M}$(M ≤ N), where m is the number of words. The model uses all input characters for self-attention calculations to obtain a character-level attention score matrix A_w. The A_w can be obtained by:

$$A_w = \text{softmax}\frac{(KW_1)(QW_2)^T}{\sqrt{d}} \quad (1)$$

Where $W_1 \in \mathbb{R}^{d\times d}$ and $W_2 \in \mathbb{R}^{d\times d}$. K and W are both equal to the collective representation in the last layer of the Chinese pre-training model. Moreover, we write matrix A as a collection of strings:

$$A_w = \text{A} = [\vec{a}_W^1, \vec{a}_W^2, \dots, \vec{a}_W^n]$$

Where $\vec{a}_{W_1}^i \in \mathbb{R}^d$, which is the i-th row vector of A_w. If $\{W_m\}_{m=1}^{M} = [\{s_1, s_2, s_3\}, \{s_4\}, \dots, \{s_n\}]$, then

$$B_m = [\{\vec{a}_W^1, \vec{a}_W^2, \vec{a}_W^3\}, \{\vec{a}_W^4\}, \dots, \{\vec{a}_W^n\}] = [\vec{b}_1, \vec{b}_2, \dots, \vec{b}_m]$$

Where the B_m is the vector representation of $\{W_m\}_{m=1}^{M}$. Subsequently, the model normalizes the vectors of each group to allow each word to obtain a row vector code representation $\vec{c}_W^i$. For example, the word "水稻" contains the two characters "水" and "稻". Therefore, the representation vectors of the two characters "水" and "稻" are divided into a group. After normalization, the representation vector of "水稻" is obtained. The model uses a combination of average pooling and maximum pooling to normalize word vectors. The detailed process can be summarized as:

$$\vec{c}_W^i = \gamma\text{Maxpooling}(\vec{b}_i) + (1-\gamma)\text{Avepooling}(\vec{b}_i) \quad (2)$$

Where γ is a learnable parameter. However, this kind of representation that only considers the current vocabulary information will cause the matching results of the vocabulary to fail to restore the information of the original vocabulary. For example, the representation vectors obtained by "稻" in the words "稻瘟病" and "水稻" are the same. In addition, the position information of the characters is not considered. Given the requirements of the algorithm, this study uses the "BMES" expression strategy to represent the character position information afterword segmentation, where B represents the beginning of the entity, M represents the middle part of the word, E represents the end of the word, and S represents the word represented by a single character. In "稻瘟病" and "水稻," "稻" is labeled as B in the former and M in the latter. Hence, "稻瘟病" and "水稻" are assigned to different sets. Finally, the words in each set are weighted to obtain the representation vectors of "稻" in different positions:

$$\vec{c}_i' = \frac{1}{(D+E)}\sum(d(\vec{c}_W^i)\vec{c}_W^i + e(\vec{c}_W^j)\vec{c}_W^j) \quad (3)$$

Where D is the number of words in set 1, E is the number of words in set 2, and $d(\vec{c}_W^i)$ and $e(\vec{c}_W^j)$ are the word frequencies in the corresponding set. To make the calculated vector dimension consistent with the initial vector, the model uses up sampling to integrate the two vectors: $\vec{c}_i'$ is assigned to the position of all characters contained in the word in the original string. We use A_{nw} to represent the aligned attention matrix:

$$A_{nw} = [\{\vec{a}_{nw}^1, \vec{a}_{nw}^1, \vec{a}_{nw}^1\}, \{\vec{a}_{nw}^2\}, \dots, \{\vec{a}_{nw}^m\}]$$

To fuse the contextual information, the model calculates the attention matrix and the original pre-training vector representation matrix P:

$$Sco = A_{nw}PW_n \quad (4)$$

At the same time, the model uses a multi-head attention mechanism to obtain different contextual information Sco_i. The model integrates different context information with the output of the word segmentation tool:

$$\bar{H} = Concat(Sco_1, Sco_2, \ldots, Sco_k)\, W_0 \quad (5)$$

Where k is the number of attention matrices. The difference in word segmentation granularity, segmentation error, and knowledge of word segmentation tools may lead to unsatisfactory model performance. Therefore, the model needs to integrate the information of multiple word segmentation tools. We assume that the integration of the vocabulary information obtained by each word segmentation tool and character representation vector is expressed as $\bar{H}$. Then, the final encoding output can be obtained by:

$$\tilde{H} = \sum_{m=1}^{M} \tanh\left(\bar{H} W_g\right) \quad (6)$$

Where W_g is a learnable weight parameter and $\tilde{H}$ is the final representation of the final character vector representations.

Some data texts have a lot of nested nouns. For example, “水稻纹曲病” is composed of two entities: “水稻” and “纹曲病.” These nested nouns will affect the results of the word segmentation tool. Given that various word segmentation tools have different recognition granularity, this result will have different semantics. Here, we have listed the word segmentation results of six commonly used Chinese word segmentation tools for “咪鲜·杀螟丹可湿性粉剂,” “水稻纹曲病,” and “稻瘟病.” The results are shown in **Table 3**.

Table 3. Comparison of segmentation effects of six commonly used Chinese word segmentation tools

Tool	咪鲜·杀螟丹可湿性粉剂	水稻纹曲病	稻瘟病
Jieba[3]	咪鲜/ ·/ 杀/ 螟丹/ 可湿性/ 粉剂	水稻/ 纹曲病	稻瘟病
Thulac [33]	咪鲜/ ·/ 杀螟丹可湿性/ 粉剂	水稻/ 纹曲病	稻瘟/ 病
LAC [34]	咪鲜·杀螟丹/ 可湿性粉剂	水稻纹曲病	稻瘟病
HanLP[4]	咪/ 鲜/ ·/ 杀/ 螟/ 丹/ 可湿性/ 粉剂	水稻/ 纹/ 曲/ 病	稻瘟病
Pkuseg [35]	咪鲜·杀螟丹/ 可湿性/ 粉剂	水稻/ 纹曲病	稻瘟病
SnowNLP[5]	咪/ 鲜/ ·/ 杀/ 螟丹/ 可湿性/ 粉/ 剂	水稻/ 纹/ 曲/ 病	稻/ 瘟/ 病

As shown in **Table 4**, LAC is the best for boundary recognition of test nouns, followed by Jieba and Pkuseg, HanLP, and SnowNLP, and Thulac, respectively. Here, we use a judgment mechanism to optimize the model. After the data text is entered, the model will count the entity nouns. If the result obtained by the word segmentation tool conflicts with the boundary of the marked entity noun, then the result is discarded. For example, if the text containing the disease

[3] https://github.com/fxsjy/jieba
[4] https://github.com/hankcs/HanLP/
[5] https://github.com/isnowfy/snownlp

term “水稻粒瘟病” (rice grain blast) is divided into “水稻粒” and “瘟病”, the result will be discarded. If it is divided into “水稻” and “粒瘟病”, the two parts will be combined to form the final result: “水稻粒瘟病.”

3.2 MM-CNN Layer

Given that the short text has strong local information association and less context information, the ability of the model to extract local information must be strengthened. Guo et al. [36] proposed a multi-core CNN to extract local context information. However, this method uses all the information extracted by CNN equally. Various sizes of CNN extract different information, which has different weights. This variation may introduce noise. Therefore, this study improves the existing work of Guo et al. and designs an MM-CNN layer enhancement model with multiple sizes and different weights to extract local information in a short text. The structure of MM-CNN is shown in **Fig. 5**.

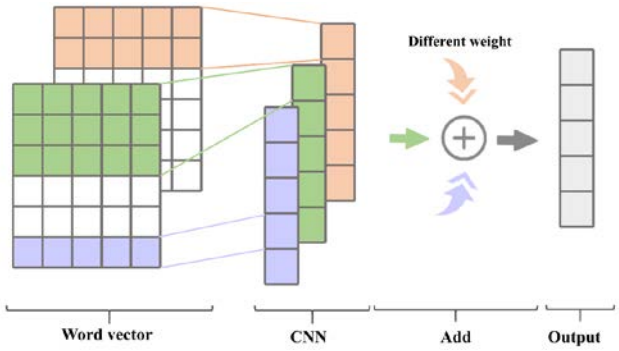


Fig. 5. The structure of the MM-CNN

The output result via the hybrid embedding layer is $S'_n = [\vec{s}'_1, \vec{s}'_2, \dots, \vec{s}'_n]$. $\vec{s}'_i$ is the vector of the ith character in the sentence, n is the length of the sentence filtered with $w \epsilon \mathbb{R}^{f \times e}$ as convolution, f is the length of each convolution, and e is the dimension of the embedding vectors. Given that CNNs of different sizes have varying abilities to extract information, this research provides different weights to the information captured by various convolution kernels. Then, the local context information of the character at position j processed by different convolution kernels is as follows:

$$c_j = Linear(H_{[i:i+j-1]} \cdot w + b) \times \frac{1}{|c_j|} \tag{7}$$

Where $|c_j|$ is the size of the convolution kernel and b is the bias weight. In this work, $Linear$ is used as the activation function of the convolution layer to retain the detailed information extracted by the convolution kernel to the greatest extent. The final output of this layer is the sum of convolutions of different sizes:

$$C_{out} = Add(c_1, c_2, \dots, c_k) \tag{8}$$

4. Results

On the basis of CCDIP, AFQMC, BQ corpus, and LCQMC datasets, this study compares AgrCQS with some classical baseline models in the field of text matching models. The comparison models involved in the comparison are mainly divided into two types: a models based on text classification and models based on deep semantic matching. The models based

on the text classification are TextCNN, TextDPCNN, BERT, XLNet, ALBERTA and RoBERTa. Meanwhile, the models based on deep semantic matching are ABCNN, RE2, ESIM, BIMPM, and DIIN.

4.1 Evaluation Index and Experiment Setup

To verify the generalization and accuracy of AgrCQS, the performance experiments of AgrCQS and different models were carried out on datasets of different fields. All experiments are based on the Keras framework, and the server platform is configured as follows: Intel(R)Core(TM)i7-9700K CPU @ 3.60GHz, 32G running memory, and Geforce GTX 1660 Ti with 6G memory. In this study, four evaluation indexes are used to evaluate the performance of the model, namely, precision (P), recall (R), F1 score (F_1), and accuracy (ACC).

To drive the model to capture more context information, this study set the sentence length to be greater than 95% as the maximum length of the embedding layer. Some super parameters are shown in **Table 4**.

Table 4. Statistics of some super parameters

Layers	Hyper-Parameters	Value
Embedding	Dimension of embedding	628
Dropout	Rate	0.2
MM-CNN	Window size	2,4,6
	Filters	140
	Stride	1
	Learning rate	0.001
BiLSTM-1	Hidden size	157
	Layer	1
Self-attention	Unit	300
BiLSTM-2	Hidden size	128
	Layer	1
Epoch	Number	20
Max Length	Number	31

4.2 Experimental Results on CCDIP

In this research, experiments are carried out on the self-built CCDIP to verify the effectiveness of AgrCQS. AgrCQS is compared with commonly used models, such as BIMPM, RE2, ABCNN, ESIM, DIIN, ALBERT, RoBERTa, BERT, XLNet, and the method proposed in this study. The experimental results are shown in **Table 5**. AgrCQS achieves good results on the CCDIP. In addition, the accuracy is 93.92%, which is the same as RoBERTa and was higher than other comparable models. This outcome verifies the effectiveness of the proposed method. However, these improvements are lower than those in AgrCQS. Although the effect of AgrCQS is slightly lower than that of RoBERTa, the number of parameters required by AgrCQS (14.3M) is far less than that of RoBERTa (102.3M). At the same time, this study verifies the time used by AgrCQS and RoBERTa to test 1020 pieces of data. The results demonstrate that the time used by AgrCQS is shorter, which is 0.51 seconds, than RoBERTa's 2.10 seconds. Therefore, the applicability of AgrCQS to the CCDIP is better than that of other models.

Table 5. Comparison of effects of various models on CCDIP

Model	P(%)	R(%)	F_1(%)	Acc(%)
RE2	89.61	84.01	86.72	86.27
ABCNN	84.12	90.51	87.2	87.65
ALBERT	91.18	89.42	90.29	90.20
RoBERTa	93.92	94.11	94.01	94.02
DIIN	89.80	89.63	89.72	89.71
XLNet	90.78	92.6	91.68	91.76
BERT	92.94	93.31	93.12	93.14
ESIM	90.98	88.21	89.58	89.41
AgrCQS	92.16	95.53	93.81	93.92

4.3 Experimental Results on Public Datasets

To verify the generalization and stability of AgrCQS, comparative experiments are carried out on the open datasets of AFQMC, BQ corpus and LCQMC. The experimental results are shown in **Table 6**. AgrCQS achieves good accuracy on three public datasets: AFQMC, BQ corpus, and LCQMC, which are 74.42%, 86.35%, and 83.05%, respectively. AgrCQS takes full account of multi-scale local context features and uses a self-attention mechanism to enhance the ability to extract sentence context further and to maximize the performance of the model. Therefore, the above experimental results reveal that the AgrCQS is not only able to measure the similarity of question corpus in the CCDIP but also has a good effect on data sets in different fields.

Table 6. Comparison of discriminant effects of different models on public datasets

Model	AFQMC	LCQMC	BQ corpus
	Acc(%)	Acc(%)	Acc(%)
TextCNN	67.81	74.90	68.52
TextDPCNN	69.83	83.70	75.08
BIMPM	68.12	83.59	81.85
XLNet	70.5	86.94	84.1
RE2	69.83	83.91	79.72
DIIN	72.13	83.71	80.12
ABCNN	64.07	82.81	78.87
ALBERT	75.6	85.22	82.2
BERT	72.09	87.14	84.2
ESIM	70.06	83.85	81.7
RoBERTa	74.02	87.26	84.6
AgrCQS	74.42	86.35	83.05

5. Discussion

5.1 Effectiveness of Hybrid Embedding Layer

To verify the effectiveness of the hybrid embedding layer, experiments are carried out on RE2, BIMPM, ABCNN, and ESIM. Word2 signifies using word2vec as an embedding layer, Emb (embedding) means using a hybrid embedding layer. As shown in **Fig. 7**, the results of the model using hybrid embedded information on CADIA-QS are significantly higher than those using word2vec. Moreover, the effect of the hybrid embedded layer is particularly significant on BIMPM, in which the accuracy increased by 5.59%. The hybrid embedding layer can use

a variety of word segmentation information, which not only enhances the model's ability to recognize the word boundary but also identifies the lexical word with different granularity. At the same time, the global lexical information embedding can summarize the lexical information of various granularity into the embedding layer, which improves the model's carrying ability of comprehensive lexical information. Therefore, the hybrid embedding layer is helpful to improve the model's ability to measure the similarity of short texts.

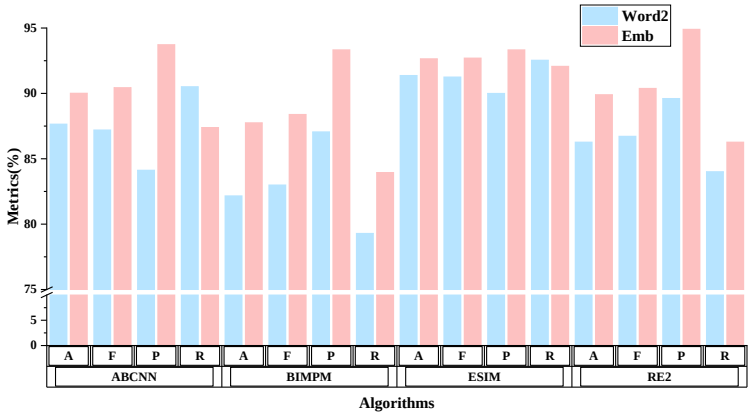


Fig. 7. Results of different model with the hybrid embedding layer on CCDIP

5.2 Impact of MM-CNN

To verify the effectiveness of the MM-CNN, experiments are carried out on the ESIM, AgrCQS-NCN (without MM-CNN), AgrCQS-CN (only MM-CNN with weight is used), AgrCQS-NWCN(only MM-CNN without weight is used), and AgrCQS on the CCDIP. The experimental results are shown in **Fig. 8**. Compared with the baseline model ESIM, the accuracy of AgrCQS-CN is improved by 3.24%, which is lower than that of the AgrCQS but better than that of AgrCQS-NWCN (91.96%). A larger window size may introduce some noise, but its weighting can weaken the effect of noise captured by the window and make use of more effective information. Although the AgrCQS-NCN is stronger than the baseline model ESIM, the accuracy is 0.49% lower than the AgrCQS. The MM-CNN used in this study can extract the local information of the input corpus at multiple scales. Through information fusion, the ability of the model to capture the local context information is enhanced, and more effective information can be used. Therefore, using the MM-CNN layer helps to improve the model's ability to measure the similarity of short text.

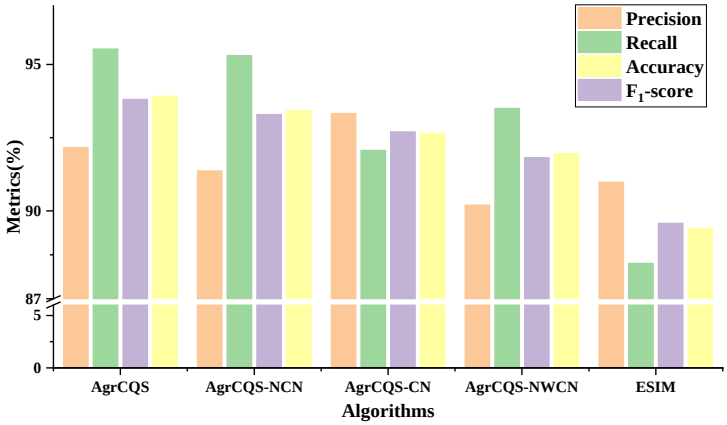
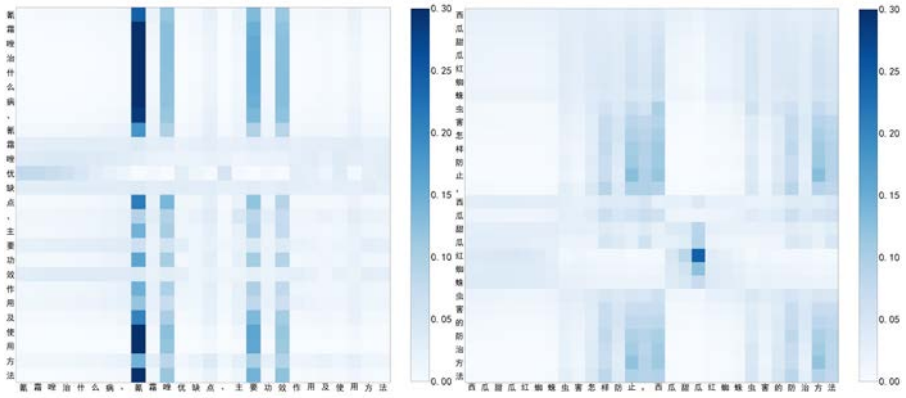


Fig. 8. Results of different model with the MM-CNN layer on CCDIP

5.3 The Influence of Self-attention Mechanism

As shown in **Table 5**, compared with the baseline model ESIM, the ESIM with self-attention mechanism further improves the ability of the model to measure the sentence similarity in the CCDIP, with the precision increasing by 1.57%. In addition, the accuracy increases by 3.04%, the recall increases by 4.16%, and the F_1 increases by 2.88%. After adding a self-attention mechanism, the judgment ability of the model improves, and it can better distinguish different types of sentences.

This paper shows the effect of the model on the attention mechanism of PHAR-IN and CON-PEST, as shown in **Fig. 9**. The results demonstrate that the self-attention mechanism can capture sentence context information from input characters and assign different weights to different characters. **Fig. 9 (a)** shows that the model can extract the relevant dependency information, such as "治什么病," "功效," "作用," and "使用方法" for the word "氰霜唑." Meanwhile, **Fig. 9 (a)** indicates that the separator "、" also plays an important role in understanding the sentence meaning of the model. The results in **Fig. 9 (b)** demonstrate that the model can recognize the core character "虫害如何防治" when processing the CON-PEST corpus, which indicates that the model has better ability to extract the core sentence meaning of input sentences. Moreover, the model can provide higher weight to the key characters in sentences. Therefore, the self-attention mechanism has a good effect on the measurement of sentence similarity in the CCDIP.



(a) (b)

Fig. 9. The weight of attention mechanism on PHAR-IN (a) and CON-PEST (b)

5.4 Quality Analysis of CCDIP

To verify the quality of the CCDIP, this study selects a traditional method and four deep learning models to evaluate the data set. The results are shown in **Table 7**. The final results are the average values obtained by using the 10-fold cross-validation method.

The method based on Cilin models sentences according to the existing vocabulary information in a word forest corpus. Finally, the approach uses Euclidean distance to calculate the similarity between sentences. This method cannot model sentences according to the existing dependency between characters. Hence, the effect is poor, and the accuracy is only 54.08%. Therefore, the CCDIP constructed in this research is unable to use simple word semantic representation to calculate the similarity between sentences. Meanwhile, four deep learning models, TextCNN, TextDPCNN, BIMPM, and ESIM are used to model the features in the corpus. The results of four deep learning models on the corpus are good, especially

BIMPM and ESIM, which use the Siamese network structure. The accuracy of the two models is 87.84% and 88.75%, respectively. To analyze the difficulties in the database, this study lists some wrong judgments, as shown in **Table 8**. When measuring the similarity between very short sentences and long sentences, some models tend to misunderstand the meaning of very short sentences, as shown in Group 2. Meanwhile, each model is prone to making mistakes in judging the semantics of multiple intention questions, as shown in Group 3. “核桃叶片早衰发生原因有哪些” belongs to the CAU-DS category. However, the question “葡萄低产是什么原因，该如何解决” belongs to CAU-PRE. Although this sentence has two simultaneous intentions, some models tend to ignore one of the intentions. Therefore, the corpus presented in this study is effective and can be used for the measurement of similarities in short texts.

Table 7. Evaluation results of each model on QSM-CCD&IP

Model	P(%)	R(%)	F_1(%)	Acc(%)
Cilin	65.73	53.28	58.83	54.08
TextCNN	71.02	77.72	73.81	74.96
TextDPCNN	80.94	80.19	79.95	80.04
BiMPM	88.63	87.26	87.94	87.84
ESIM	88.24	89.21	88.67	88.75

Table 8. Error discrimination of partial AgrCQS on CCDIP

Model	P(%)	R(%)	F_1(%)	Acc(%)
1	幸福树叶子发黄的原因 What is the reason for the yellow leaves of radermachera sinica?	CAU-DS	Yes	False
	稻瘟病发病的主要原因是什么？ What is the main cause of the rice blast?	CAU-DS		
2	矮稻的病原 What is the pathogen of dwarf rice?	PATH-CD	False	Yes
	松材线虫病是一种啥疾病？对植物有啥危害？ What kind of disease is SteineretBuhrer? What harm does it do to plants?	DIS-IN		
3	核桃叶片早衰发生原因有哪些 What are the causes of presenility pathogenesis of Walnut Leaves?	CAU-DS	False	Yes
	葡萄低产是什么原因，该如何解决？ What are the reasons for the low yield of grapes and how to solve them?	CAU-PRE		
4	扁豆不出苗怎么办，有哪些方法措施 How do lentils not sprout to do? What are the measures?	PRE-DS	False	Yes
	芹菜腐烂病怎么回事，芹菜腐烂病的防治方法 What is the cause of the soft rot of celery? What are the control methods of soft rot of celery?	CAU-PRE		

6. Conclusions

In this research, three studies have been carried out to solve the problems of missing databases and methods in the field of QSM-CCD&IP.

First, we construct a similarity matching corpus CCDIP involving 13 categories of CCQ. The corpus contains 10559 labeled sentences and 5100 question matching pairs. Compared with the data sets of other fields, CCDIP has richer domain vocabulary and ambiguous lexical boundary features. The experimental results on the CCDIP using the 10-fold cross-validation method reveal that the best accuracy is 88.75% (ESIM), which indicates that the dataset has certain challenges and verifies the rationality of the model.

Secondly, AgrCQS, a deep learning model for the QSM-CCD&IP is proposed to solve the low-level network error propagation caused by the difficulty of recognizing word boundaries and the insufficient ability to extract local information from a short text. The experimental results reveal that the hybrid embedding layer of the AgrCQS proposed in this work can better solve the problem regarding the model and the conventional character embedding layer not being able to make full use of the lexical information. The MM-CNN with a multi-scale proposed in this research enhances the ability of the model to extract local information and further enriches the semantic information. Moreover, the use of the self-attention mechanism in the model improves the ability of the BiLSTM to capture long-distance dependency and strengthens the ability of the model to extract core semantics. Experiments on the CCDIP demonstrate that the model can effectively distinguish the similarity of the question intention with an accuracy of 93.92% and F_1 at 93.81%.

Third, this research compares the AFQMC, LCQMC, BQ corpus, and other field datasets with BERT, XLNet, and other models. The comparison achieves good results. The accuracy is 74.42%, 86.35%, 83.05%, respectively, which are higher than the BIMPM, ESIM, and other models, and equal to the BERT series models. The experimental results reveal that the recognition effect of the AgrCQS is the same as that of the BERT series model in the case of fewer computing resources, thus verifying the generalization of the model.

Furthermore, the method module proposed in this study can be disassembled separately and transplanted to other models, which can provide a reference for the design of the text-similarity measurement model. In the future, the author aims to overcome the following three challenges to improve QSM-CCD&IP research.

(1) The author will expand the number of data sets to achieve more different sentence patterns and different types of sentences meaning discrimination.

(2) This study used the question itself as basis to judge its intention similarity with other questions. In the follow-up research, we will attempt to assess the intention similarity of questions based on the similarity of the answers.

(3) Future research must design effective discriminant rules, such as voting discrimination mechanisms, for sentences with similar sentence patterns but different intentions.

References

[1] R. B. Zhang, Q. C. Zhang, Y. Zhou, Y. Pan, M. Zhu, Y. Qi and X. X. Sun, “Occurrence and control measures of rice blast in Jianhu County in 2019,” *Anhui Agricultural Science Bulletin*, vol. 26, no. 24, pp. 127-128, December 2020. Article (CrossRef Link)

[2] L. Javier, L. P. F. Javier, E. G. Borja, N. I. Javier and Z. S. F. Javier, “Agricultural recommendation system for crop protection,” *Computers and Electronics in Agriculture*, vol. 152, pp. 82-89, September 2018. Article (CrossRef Link)

[3] N. Ghasemi and S. Momtazi, “Neural text similarity of user reviews for improving collaborative filtering recommender systems,” *Electronic Commerce Research and Applications*, vol. 45, pp. 1567-4223, January–February 2021. Article (CrossRef Link)

[4] Y. Z. Zhang, D. W. Song, P. Zhang, X. Li and P. P. Wang, "A quantum-inspired sentiment representation model for twitter sentiment analysis," *Applied Intelligence*, vol. 49, pp. 3093–3108, March 2019. Article (CrossRef Link)
[5] S. Cox, X. L. Dong, R. Rai, L. Christopherson, W. F. Zheng, A. Tropsha and C. Schmitt, "A semantic similarity based methodology for predicting protein-protein interactions: Evaluation with P53-interacting kinases," *Journal of Biomedical Informatics*, vol. 111, pp. 1532-0464, November 2020. Article (CrossRef Link)
[6] M. Zhao, C. C. Dong, Q. X. Dong and Y. Chen, "Question Classification of Tomato Pests and Diseases Question Answering System Based on BIGRU," *Transactions of The Chinese Society of Agricultural Machinery*, vol. 49, no. 5, pp. 271-276, October 2018. Article (CrossRef Link)
[7] M. Y. Zhang, H. R. Wu and H. J. Zhu, "Analysis of Extraction of Semantic Feature in Agricultural Question and Answer Based on Convolutional Model," *Transactions of the Chinese Society for Agricultural Machinery*, vol. 49, no. 12, pp. 203-210, May 2018. Article (CrossRef Link)
[8] N. Jin, C. J. Zhao, H. R. Wu, Y. S. Miao, S. Li and B. Z. Yang, "Classification Technology of Agricultural Questions Based on BiGRU_MulCNN," *Transactions of The Chinese Society of Agricultural Machinery*, vol. 51, no. 5, pp. 199-206, August 2020. Article (CrossRef Link)
[9] M. J. Kusner, Y. Sun, N. I. Kolkin and K. Q. Weinberger, "From word embeddings to document distances," in *Proc. of the 32nd International Conference on Machine Learning, ICML*, vol. 37, pp. 957-966, July 2015. Article (CrossRef Link)
[10] M. A. Álvarez-Carmona, M. Franco-Salvador, E. Villatoro-Tello, M. Montes-y Gómez, P. Rosso and L. Villaseñor-Pineda, "Semantically-informed distance and similarity measures for paraphrase plagiarism identification," *Journal of Intelligent & Fuzzy Systems*, vol. 34, no. 5, pp. 2983-2990, May 2018. Article (CrossRef Link)
[11] T. Kenter and M. de Rijke, "Short text similarity with word embeddings," in *Proc. of the 24th ACM International Conference on Information and Knowledge Management, CIKM*, pp. 1411-1420, October 2015. Article (CrossRef Link)
[12] Z.C. Zhang, "An improved BM25 algorithm for clinical decision support in Precision Medicine based on co-word analysis and Cuckoo Search," *BMC Medical Informatics and Decision Making*, vol. 21, no. 81, March 2021. Article (CrossRef Link)
[13] D. Suhartono and K. Khodirun, "System of Information Feedback on Archive Using Term Frequency-Inverse Document Frequency and Vector Space Model Methods," *International Journal of Informatics and Information Systems*, vol. 3, no. 1, pp. 36-42, 2020. Article (CrossRef Link)
[14] Z. Quan, Z. Wang, Y. Le, B. Yao, K. Li and J. Yin, "An Efficient Framework for Sentence Similarity Modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 853-865, April 2019. Article (CrossRef Link)
[15] A. Severyn, M. Nicosia and A. Moschitti, "Building structures from classifiers for passage reranking," in *Proc. of the 22nd ACM international conference on Information & Knowledge Management*, pp. 969-978, October 2013. Article (CrossRef Link)
[16] Y. Kim, "Convolutional Neural Networks for Sentence Classification," in *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746-1751, October 2014. Article (CrossRef Link)
[17] R. Johnson and T. Zhang, "Deep Pyramid Convolutional Neural Networks for Text Categorization," in *Proc. of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 562-570, July 2017. Article (CrossRef Link)
[18] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171- 4186, June 2019. Article (CrossRef Link)
[19] Z. L. Yang, Z. H. Dai, Y. M. Yang, J. Carbonell, R. Salakhutdinov and Q. V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," *arXiv preprint arXiv: 1906.08237*, January 2019. Article (CrossRef Link)

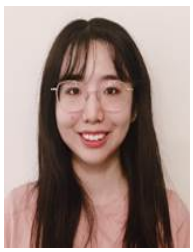
[20] Y. H. Liu, M. Ott, N. Goyal, J. F. Du, M. Joshi, D. Q. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv: 1907.11692*, July 2019. Article (CrossRef Link)
[21] Z. Z. Lan, M. D. Chen, S. Goodman, K. Gimpel, P. Sharma and R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," *arXiv preprint arXiv: 1909.11942*, September 2019. Article (CrossRef Link)
[22] W. P. Yin, H. Schütze, B. Xiang and B. W. Zhou, "ABCNN: Attention-Based Convolutional Neural Network for Modeling Sentence Pairs," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 566-567, December 2016. Article (CrossRef Link)
[23] Q. Chen, X. D. Zhu, Z. H. Ling, S. Wei, H. Jiang and D. Inkpen, "Enhanced LSTM for Natural Language Inference," in *Proc. of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 1657-1668, July 2017. Article (CrossRef Link)
[24] Z. G. Wang, W. Hamza, R. Florian, "Bilateral Multi-Perspective Matching for Natural Language Sentences," in *Proc. of the 26th International Joint Conference on Artificial Intelligence*, pp. 4144-4150, 2017. Article (CrossRef Link)
[25] R. Q. Yang, J. H. Zhang, X. Gao, F. Ji and H. Q. Chen, "Simple and Effective Text Matching with Richer Alignment Features," in *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4699-4709, July 2019. Article (CrossRef Link)
[26] X. Liu, Q. C. Chen, C. Deng, H. J. Zeng, J. Chen, D. F. Li and B. Z. Tang, "LCQMC: A Large-scale Chinese Question Matching Corpus," in *Proc. of the 27th International Conference on Computational Linguistics*, pp. 1952-1962, August 2018. Article (CrossRef Link)
[27] J. Chen, Q. C. Chen, X. Liu, H. J. Yang, D. H. Lu and B. Z. Tang, "The BQ Corpus: A Large-scale Domain-specific Chinese Corpus for Sentence Semantic Equivalence Identification," in *Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4946-4951, October-November 2018. Article (CrossRef Link)
[28] L. Xu, X. Zhang and Q. Dong, "CLUECorpus2020: A Large-scale Chinese Corpus for Pre-training Language Model," *arXiv preprint arXiv: 2003.01355*, March 2020. Article (CrossRef Link)
[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is All You Need," in *Proc. of the 31st International Conference on Neural Information Processing Systems*, pp. 6000-6010, December 2017. Article (CrossRef Link)
[30] M. T. Luong, H. Pham and C. D. Manning, "Effective Approaches to Attention-based Neural Machine Translation," in *Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing*, *Association for Computational Linguistics*, pp. 1412–1421, September 2015. Article (CrossRef Link)
[31] S. Hochreiter, J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no.8, pp. 1735–1780, November 1997. Article (CrossRef Link)
[32] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *arXiv preprint arXiv: 1301.3781*, January 2013. Article (CrossRef Link)
[33] L. Zhongguo and S. Maosong, "Punctuation as Implicit Annotations for Chinese Word Segmentation," *Computational Linguistics*, vol. 35, no. 4, pp. 505-512, December 2009. Article (CrossRef Link)
[34] Z. Y. Jiao, S. Q. Sun and K. Sun, "Chinese lexical analysis with deep Bi-GRU-CRF network," *arXiv preprint arXiv: 1807.01882*, July 2018. Article (CrossRef Link)
[35] R. X. Luo, J. J. Xu, Y. Zhang, X. C. Ren and X. Sun, "PKUSEG: A toolkit for multi-domain Chinese word segmentation," *arXiv preprint arXiv: 1906.11455*, June 2019. Article (CrossRef Link)
[36] X. C. Guo, H. Zhou, J. Su, X. Hao, Z. Tang, L. Diao and L. Li, "Chinese agricultural diseases and pests named entity recognition with multi-scale local context features and self-attention mechanism," *Computers and Electronics in Agriculture*, vol. 179, December 2020. Article (CrossRef Link)
[37] M. Mirakyan, K. Hambardzumyan and H. Khachatrian, "Natural language inference over interaction space: ICLR 2018 reproducibility report," *arXiv preprint arXiv: 1802.03198*, February 2018. Article (CrossRef Link)

HAN ZHOU received his Bachelor's Degree in Computer Science and Technology in 2015 from Shanghai University of Electric Power, Shanghai China. He is currently pursuing the Master's degree with the College of Information and Electrical Engineering, China Agricultural University, Beijing China. His research interests include: Complex Network, Knowledge Graph, Text Mining, etc.

XUCHAO GUO received his Bachelor's degree in Computer Science and his Master's degree in 2018 from Shandong Agricultural University. He is currently Ph.D. student in the College of Information and Electrical Engineering, China Agricultural University. He is mainly engaged in natural language processing and knowledge graph in agricultural fields. His research interests include: deep learning, complex network analysis, data mining, machine learning, and image processing.

CHENGQI LIU received the master's degree from China Agricultural University, Beijing, China, in 2018, where he is currently pursuing the Ph.D. degree. His research interests include image processing and deep learning for video processing.

ZHAN TAN received the M.S. degree in computer science from Shanghai Normal University. He is currently pursuing the Ph.D. degree with the College of Information and Electrical Engineering, China Agricultural University, Beijing, China. His research interests include natural language processing, text mining, and machine learning.



SHUHAN LU received her Bachelor's degree in Computer and Information Science in 2020 at the Ohio State University. From 2020, she is currently studying Master's in Health Informatics in University of Michigan, Ann Arbor. She has good experience in database creating, database managing, and data analysis. Her research interests include: data mining, database manage, machine learning and image processing.

LIN LI is currently a Professor and a Doctoral Supervisor with the College of Information and Electrical Engineering (CIEE), China Agricultural University. Her main research interests include knowledge engineering and machine learning.