Short Text Similarity Computing Method towards Agriculture Question and Answering Systems

Bingjie Sun, Zhichao Liang, Qingtian Zeng+, Hua Zhao, Weijian Ni, Hua Duan
College of Information Science and Engineering, Shandong University of Science and Technology
Qingdao, Shandong, China
+Corresponding author
E-mail: qtzeng@163.com

***Abstract*—Text similarity computing is the core issue that question-answering system needs to solve. It is mainly used to filter out the existed problems which are similar to the user's questions from database. Because of the low recall of domain keywords in domain text similarity computing based on traditional semantic dictionary, this paper proposed a short text similarity computing method in the field of agriculture based on the extended version of <<Tongyicicilin>> which referred to as <<CiLin>>. This paper propose to consider both the similarity and correlation when calculate the words' final similarity. The experimental results show that the proposed short text similarity computing method resolve the problem of the low recall of domain words in traditional semantic dictionary well, and improve the similarity calculation performance of high relevant keywords greatly.**

Keywords- Agricultural question-answering syste；Semantic dictionary；Text similarity；Similarity；Correlation

I. INTRODUCTION

Question and answering system (QA) is concerned more and more in recent years [1, 2]. In QA system, the real-time performance and efficiency of the research can be improved by communicating with users, which is the so-called user interactive system [3, 4, 5]. Agriculture interactive QA relies on the participation of the users to find a proper way to solve the problems user encountered in the process of agricultural production. And it provides an efficient and convenient way to farm users to solve practical problems during the agricultural production.

The main problem agriculture interactive QA need to solve is how to obtain the answers users expected from the large-scale agriculture domain short texts. Where, one of the core questions is to retrieve the existed questions which are similar to the user's question by computing the similarity between the questions. There exists several typical text similarity calculation methods, which include text similarity calculation based on the vector space model [6], text similarity calculation based on edit-distance [7], and text similarity calculation based on semantic dictionaries [6,8]. Text similarity calculation method based on vector space model is applicable to large-scale text. However the texts in the QA of this paper are shorter, and contain fewer words. This method only considers the word similarity, and will face the problem of data sparse, which can hardly ensure the accuracy of the retrieval. Text similarity calculation method based on edit-distance only considers the issue of the composition of the text itself, and ignores the semantic information of the text as well as internal words. So, the performance of this method is not acceptable. Text similarity calculation based on the semantic dictionaries considers the correlations between the words, extends the semantic of the words in the text and fits short texts well. However, because of the insufficiency of the included domain words, traditional similarity calculation methods based on the semantic dictionaries can hardly satisfy the need of specific domain QA. In the result, the recall of the domain words is low and the research performance can't satisfy the users' demand. So, on the basis of effective semantic extensions to the agricultural short texts, this paper proposed a final similarity calculation method which considered word similarity and word correlation together. Experimental results show that this method can overcome the defects of semantic in recall.

II. SHORT TEXT SIMILARITY CALCULATION PROBLEM IN AGRICULTURE QA

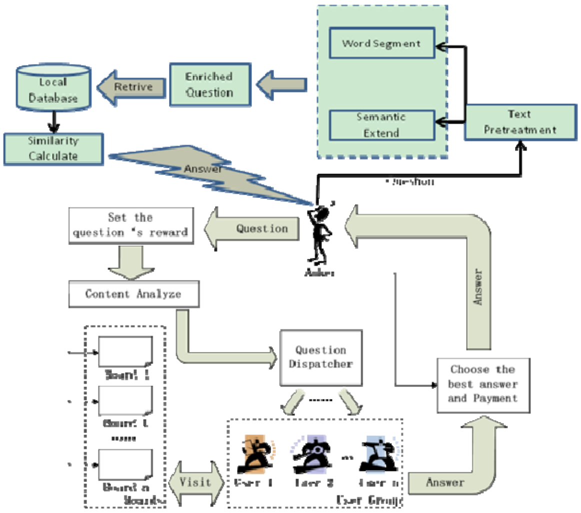

Figure 1. Architecture of Agriculture Interactive QA System

Agriculture interactive QA system uses the strategy "retrieves firstly, interacts secondly". Users can retrieve the system database to find answers for their questions. The retrieval results will be ranked and listed according to the similarity with users' questions. If users aren't satisfy with the results, they could submit their questions in the interaction process and get the results. The architecture of agriculture interactive QA system is shown in Figure 1.

During the process of the similar questions retrieval, QA system will carry out the following steps: (1) pre-process the texts of users' questions (such as, Chinese word segmentation and semantic extension) [9,10,11]; (2) retrieve the database, and calculate the similarity between the semantic extended question and texts in database; (3) return the similar questions to the users.

Agricultural domain short text similarity calculation is proposed in the context of the QA system. The experimental results proved that the similarity calculation method proposed in this paper can effectively improve the performance of QA system, and the user satisfaction.

III. Agricultural Short Text Similarity Calculation Method

The text similarity calculation is divided into two parts in this paper: text preprocessing and similarity calculation. After Chinese word segmentation in the process of text preprocessing, the retrieved keywords set and the keywords set of the text to be compared were obtained. Extend the semantic of the keywords based on the semantic dictionary and calculate the keywords' similarity. By comparing the method proposed in this paper with the method based on Hownet and the method based on <<CiLin>>, the method proposed in this paper performs better.

A. *Semantic Similarity Calculation Method*

Take advantage of Chinese word segmentation with domain dictionary, and get the keywords set from the user retrieved questions ($cluster1 = \{word_{11}, word_{12} \cdots word_{1n}\}$). Keywords set obtained from the questions are stored in the database ($cluster2 = \{word_{21}, word_{22} \cdots word_{2m}\}$). Each keyword in keywords set $cluster1$ and $cluster2$ has one or more item explanations [8]. This paper will use <<CiLin>> expanded by Harbin Institute of Technology. It contains 70,000 words and all of the words are coded. According to the layout characteristics of <<CiLin>>, the main idea of similarity calculation based on the item explanation [8] is: calculate the similarity between the item explanations according to the distance of the two item explanations.

The item explanation expression is as follows: "*Da15B02=*". It has 8 coded bits. The 1st code bit represents the "Class". The 2nd code bit represents the "Midclass". The 3rd and 4th code bits represent the "Subcalss". The 5th code bit represents the "Word group". The 6th and 7th code bits represent the "Atom word group". The 8th code bit is tail number.

The similarity calculation method of item explanation is as follows [8] $(1 \le i \le n, 1 \le j \le m)$:

(1) If the two item explanations are on the different trees,

$$similar(word_{1i}, word_{2j}) = f \tag{1}$$

(2)If the two item explanations are on the same tree,

- If the difference appears on the second branch, the coefficient is a ,

$$similar(word_{1i}, word_{2j}) = 1 \times a \times \cos[n \times \frac{\pi}{180}][\frac{n-k+1}{n}] \tag{2}$$

- If the difference appears on the third branch, the coefficient is b ,

$$similar(word_{1i}, word_{2j}) = 1 \times 1 \times b \times \cos[n \times \frac{\pi}{180}][\frac{n-k+1}{n}] \tag{3}$$

- If the difference appears on the forth branch, the coefficient is c ,

$$similar(word_{1i}, word_{2j}) = 1 \times 1 \times 1 \times c \times \cos[n \times \frac{\pi}{180}][\frac{n-k+1}{n}] \tag{4}$$

- If the difference appears on the fifth branch, the coefficient is d ,

$$similar(word_{1i}, word_{2j}) = 1 \times 1 \times 1 \times 1 \times d \times \cos[n \times \frac{\pi}{180}][\frac{n-k+1}{n}] \tag{5}$$

Where, n is the total number of branch nodes, k is the distance between the two branches. If the branches of the two item explanations are same, the similarity of the item explanations can be calculated by the tail number. If it is "=", the similarity is 1. If it is "#", the similarity is e . If it is "@", this condition will not be considered. After manual evaluation, the coefficient values in item explanation similarity calculation are as follows [8]:

a =0.65; b =0.8; c =0.9; d =0.96; e =0.5; f =0.1

The similarity of two keywords is the distance of item explanations related to the keywords. Calculate the similarity of these item explanations between each other and select the biggest value as the similarity value of the two keywords.

This paper uses bi-directional semantic similarity calculation methods. The so-called bi-directional semantic similarity calculation firstly calculates the similarity of each keyword in $cluster1$ with that in $cluster2$, then reverse the direction of calculation. Finally, the semantic similarity ($sim\_semantic$) is obtained by averaging the two similarities coming from the two directions.

$$sim\_semantic = \frac{1}{2} \times \left(\frac{\sum similar_{(cluster1 \to cluster2)}}{n} + \frac{\sum similar_{(cluster2 \to cluster1)}}{m} \right) \tag{6}$$

Where, $similar_{(cluster1 \to cluster2)}$ and $similar_{(cluster2 \to cluster1)}$ can be obtained by one of the formulas from formula (1) to formula (5).

This paper will choose the top-n most similar questions from the database to return to the users. Compared with the similarity calculation method simply rely on the keywords hit number, this method improves the precision largely.

B. *Keywords Correlation Calculation Method*

There is a problem of low recall for some domain keywords using ordinary similarity calculation method based on semantic dictionary, which is because some of the domain keywords have no item explanations in the semantic dictionary. So, the similarity of these keywords can’t be calculated. This paper mainly improves the performance of similarity calculation method of the expanded <<CiLin>>. The similarity and correlation are explained as follows.

(1) Similarly: The two words having some semantic similarity reflect the two words in the language have some substitutability. The similarity reflects the substitutability to a certain extent. [12]

(2) Correlation: It reflects the interdependent and mutual influence of the two words in semantic. [12] This kind of relationship between words is generally not mutually alternative.

Suppose there are two keywords $word_1$ and $word_2$. The keyword correlation calculation is done as follows. Take $word_1$ and $word_2$ as input receptively, and use the API provided by Baidu search engine, the result related to $word_1$ is m, and $word_2$ is n. The concurrent number of $word_1$ and $word_2$ is h. Then, the correlation of $word_1$ and $word_2$ is as follows:

$$relative(word_1, word_2) = \frac{h}{n+m-h} \tag{7}$$

Now the final similarity calculation formula is as follows:

$$Sim(word_{1i}, word_{2j}) = \alpha \times sim\_semantic(word_{1i}, word_{2j})$$
$$+(1-\alpha) \times relative(word_{1i}, word_{2j}) \tag{8}$$

Where, $sim\_semantic(word_{1i}, word_{2j})$ represents the semantic similarity of keywords, and $relative(word_{1i}, word_{2j})$ represents the correlation of keywords. α is an adjustment parameter. Its value is an experimental value related to the importance of similarity and correlation $(1 \le i \le n, 1 \le j \le m)$.

① If $relative(word_{1i}, word_{2j}) \geqq sim\_semantic(word_{1i}, word_{2j})$, α =0.6;

② If $relative(word_{1i}, word_{2j}) < sim\_semantic(word_{1i}, word_{2j})$, α =1;

③ If $sim\_semantic(word_{1i}, word_{2j})$ has no effective value， α =0.

C. Short Text Smilarity Calculation Based on Bidirectional Mapping

To calculate the similarity of two short texts, $Text1$ and $Text2$, we use bidirectional mapping method. Firstly, calculate the similarity of each word in $Text1$ with the words in $Text2$ (shown in figure 2). Then, calculate reversely.

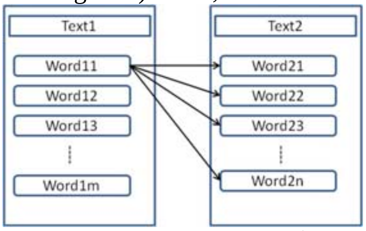

Figure 2. Similarity Calculation Model of Short Texts

The final similarity of the short texts is the average value of two directions.

$$Sim = \frac{1}{2} \times \left(\frac{\sum Sim_{Text1}}{m} + \frac{\sum Sim_{Text2}}{n} \right) \tag{9}$$

Where, m and n is the word number in $Text1$ and $Text2$. Sim_{Text1} and Sim_{Text2} can be obtained by formula (8).

IV. EXPERIMENT

To verify the effectiveness of the similarity method proposed in this paper, this paper uses Hownet and <<CiLin>> as tools, and uses domain texts as experimental data. Compare and analyze the performance of the method proposed in this paper, method based on Hownet and <<CiLin>>.This paper uses some keywords that don’t contain in the semantic dictionary or poor performance high relative keywords in general method as experimental data. Experimental results are shown in Table 1 (“?” in Table 1 represents the similarity of relevant keywords can’t be calculated.).

TABLE I. PERFORMANCE COMPARISON OF SIMILARITY CALCULATION

Group	Tuple	Method of this paper	Hownet	Method based on <<CiLin>>
1	(night willow herb, tuberose)	1	1.0	1
2	(banana, sweet banana)	1	0.01	1
3	(doctor, disease)	0.427754	0.1077581395	0.1
4	(temperature, humidity)	0.894517	0.7222222	0.8994517
5	(teacher, school)	0.346695	0.1995789474	0.1
6	(sow, tillage)	0.646632	0.104827586	0.646632
7	(urea, fertilizer)	0.573283	0.444444	0.573283
8	(plant, breed)	0.554256	0.4444445	0.55426
9	(hectare, area)	0.508068	0.044444	0.508068
10	(far, distance)	0.1	0.044444	0.1
11	(veterinarian, disease)	0.1	0.10775814	0.1
12	(wheat, plant diseases and insect pests)	0.1	0.121936842	0.1
13	(pig, culture)	0.1	0.044444	0.1
14	(oviposition, season)	0.1	0.04088889	0.1
15	(PH value, alkalinity acidity)	0.041168	0.01	?
16	(foot-and-mouth disease, guard against)	0.002456	0.01	?
17	(yellow leaf disease, prevention and cure)	0.000762	0.01	?
18	(apple, yellow leaf disease)	0.000989	0.01	?

Figure 3 shows the differences between the three methods intuitively.

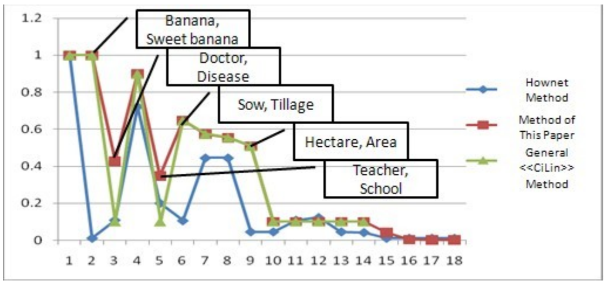

Figure 3. Methods Comparison on Data

In the case of most of the tuples' similarity results are close to each other, individual tuple will have a large difference in the calculation performance. Similarity and correlation are two aspects that determine the final similarity. For example, “Doctor” and “Disease” are not similar (The similarity of the two words is 0.1). However, their correlation is high (Its value is above 0.9). The correlation is not well reflected in Hownet and <<CiLin>>. In addition, this paper deals with the low recall of some keywords in <<CiLin>> (tuple 14 to 18). From figure 2 we can see the method in this paper performs better in keyword similarity calculation

V. Conclusions and Future Work

The contributions of this paper are as follows:

- Establish a large-scale agricultural domain word segmentation dictionary. Improve the precision of Chinese word segmentation in agricultural domain short text.
- Use the extended <<CiLin>> to add semantic information of keywords and overcome the data sparseness problem in short texts.
- This paper proposed a method to calculate the final similarity between the keywords using both similarity and correlation. It improved the calculation performance.

The method in this paper has a reference for the semantic extension and similarity calculation of both Chinese and English. Considering the correlation slows down the calculation speed. Future work can be extended to other semantic dictionaries and improve the similarity calculation efficiency.

Acknowledgment

This paper is supported partly by the NSFC (61170079 and 61202152); Sci. & Tech. Development Fund of Shandong Province of China (2010GSF10811); Specialized Research Fund for the Doctoral Program of Higher Education of China (20103718110007) and Special Fund for Fast Sharing of Science Paper in Net Era by CSTD, Sci. & Tech. Development Fund of Qingdao(10-3-3-32-nsh), Excellent Young Scientist Foundation of Shandong Province of China (BS2010DX009); China Postdoctoral Science Foundation (2011M501155); and Natural Science Foundation for Distinguished Young Scholars of Shandong and SDUST (JQ200816 and 2010KYJQ101).

References

[1] Li Liu, Qingtian Zeng. State-of-the-art of the Question-answering System, Journal of Shandong University of Science and Technology, Volume 26th issue 4th, 73-76, 2007.10.

[2] Qingtian Zeng, Zhongying Zhao, Yongquan Liang. Course Ontology-based User's Knowledge Requirement Acquisition from Behaviors within E-Learning Systems, Computers & Education, Volume 53， Issue 3 (November 2009) 809–818.

[3] Dawei Hu, Wei Chen, Qingtian Zeng, Tianyong Hao, Feng Min, Liu Wenyin, Using a User-interactive QA System for Personalized e-Learning, The International Journal of Distance Education Technologies, 6(3), 1-22, July-September 2008.

[4] Tianyong Hao, Dawei Hu, Liu Wenyin, Qingtian Zeng, Semantic Patterns for User-Interactive Question Answering, Concurrency and Computation: Practice and Experience, v 20, n 7, May, 2008, p 783-799.

[5] Chen, Wei; Zeng, Qingtian; Wenyin, Liu; Hao, Tianyong, A user reputation model for a user-interactive question answering system, Concurrency Computation Practice and Experience, v 19, n 15, October, 2007, p 2091-2103.

[6] Wanpeng Song. The use of Short text similarity computing in user interactive answering system, China Science and Technology University Doctoral Thesis, 19-51, 2010.

[7] Shengjun Ji. Sentence Similarity Computing Based on Levenshteindistance Algorithm, Computer Knowledge and Technology，2177-2178，2009.

[8] Jiule Tian, Wei Zhao. Words Similarity Calculation Method Based On TongyiciCiLin. Journal of Jilin University. 603-605，2010.

[9] Rohini Srihari, Wei Li. Information Extraction Supported Question Answering, Proceedings of TREC-8. 1999.

[10] Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Michael Junk, Chin-Yew Lin. Question Answering in Webclopedia. TREC 2000.

[11] Sanda M Harabagiu, Vinay Chaudhri. Mining answers from text and knowledge bases. AAAI Spring Sympo-sium Series, American, 2002.

[12] http://www.alibado.com/exp/detail-w1217727-e502449-p1.htm